Philosophy of Science Assoc. 23rd Biennial Mtg

San Diego, CA

Version: 13 November 2012

PhilSci ARCHIVE

Table of Contents

Catherine Allamel-Raffin

University of Strasbourg (France)

catherine.allamelraffin@unistra.fr

From Intersubjectivity to Interinstrumentality. The Example of Surface Science.

Abstract: My aim is to show how a strategy used in the experimental sciences, which I name “interinstrumentality”, can minimize the role of sociological factors when one tries to understand how the debates about the interpretation of data come to an end. To defend this view, two examples are presented. The first is historical – the invention of the Scanning Tunneling Microscope (STM) – and the second is collected during an ethnographic study in a surface science laboratory. I would like to emphasize that interinstrumentality contributes to objectivity of the experimental results and constitutes a part of it as well as intersubjectivity.

1. Introduction

Among the methodological requirements to which all scientific work is submitted, one of the first is the objectivity of processes and results. But what is objectivity? And how should one use the notion in a meta-epistemological perspective? If one thinks along with Daston (1992) that the term is unclear, it is probably because it is an umbrella-term. Let us just say that "absolute objectivity" (Megill, 1994) is based on the belief that the reality could be described literally, weeding out most of the subjectivity's effects. Such a definition constitutes more a problem than a solution. If the aim of our study is to take into account the day-to-day practices of laboratory, perhaps it would be better to agree with Putnam when he writes (2003, 142): "In scientific practice, the questions about objectivity do not concern metaphysics." Thus, objectivity has to be conceived as a *continuum*. Putnam asserts: "If we consider our statements as based on a *continuum* (…), the statements which eminently depends on our interests, on our points of view and on our idiosyncratic characteristics, are the subjective extremity of the *continuum*, whereas the statements become more objective and could pretend to be more true as they less rely on idiosyncratic points of view or on personal interests" (2003, 141).

Usually, objectivity is related to intersubjectivity. Even if contemporary philosophers of science admit that intersubjectivity is hard to define, this notion permits to emphasize on the collective dimension of the scientific activity. To constitute intersubjectivity, a communication between subjects (ideally interchangeable subjects) is

needed, their purpose is to get an assessment of their process and the evaluation of the results provided by each member. Such a definition underlines the fact that we have no more confidence in individual reason as a way to establish absolute truths. The scientific intelligence is now conceived as distributed. Besides, this definition of intersubjectivity has the advantage to avoid the *a priori* issued of the fundationalists positions (divine guarantee, first principles…).

Even if this definition of objectivity presents many advantages, we have to keep its limits in mind. Indeed, the term "intersubjectivity" is not consensual: perception's agreement in a phenomenological perspective; critical thinking of the researchers which allow them to eliminate false theories, and determine which of the remaining theories is the best available one; collective confidence attached to discipline's standards. In all definitions, for philosophers as for sociologists or for historians of science, a central feature is the concordance of the individual points of view. And none could reasonably dispute such assertion: the subjects are, ultimately, the ones who assign the signification to data, and who decide how they will validate the knowledge as true and justified. But this assertion could be expressed more or less radically. We could consider that the construction of consensus in the scientific community may rely on determining factors such as psychological characteristics, know-how, tacit knowledge of scientists themselves. This consensus could also rely on the social contexts in which those researches appeared. In that matter of case, the natural world's phenomena, which are studied, play only a limited role in the elaboration of scientific knowledge. This is the point of view of cognitive relativism.

In the rest of my paper, I will focus on a contemporary version of this cognitive relativism: the relativistic Sociology of Science lead by Harry M. Collins[1]. This relativism is assumed as a fundamental injunction for the methodology of empirical studies. The methodology at work in this kind of sociology relies on observation of day-to-day practices as they occur in laboratory (either with an ethnographic approach or with historical studies). According to this approach, such observations permit to produce a non reductivist description of the scientific work.

My study also tries to pay attention to the day-to-day activities in scientific laboratory, but it gives results which contradict previous conclusions promoted by the relativistic sociology of science, as I will show it in the following pages. The attention given to the practices elicits strategies, which aim to strengthen the objectivity of experimental results and, by the way, refutes the fact that objectivity only relies on social consensus. Interinstrumentality is the name I give to the strategy described in this paper.

In a first part, I specify briefly the thesis on which I opposite: Collins' Empirical Programme of Relativism**.** In a second part, I use an historical example - the invention of STM– to carry off my critic. This illustration permits to identify the role of interinstrumentality, when the validity of a new instrument is settled. In a third part, I will underline that such a strategy does not characterize only exceptional events (as the invention of a new scientific instrument). It could also appear in the day-to-day activities

of surface science laboratory, as the second example, borrowed from my ethnographic studies[2], will underline it.

2. Collins's Empirical Programme of Relativism

Collins brings out the stages of an "Empirical Programme of Relativism" (EPOR), since his synthetic article of 1981. The stages of the programme (stages 1 and 2 of EPOR) aim:

- to demonstrate that experimental data are submitted to a so-called "interpretative flexibility";
- to clarify the processes which give an end to the debate between competing interpretations.

2.1. The Interpretative Flexibility of Data

In Collins' studies, the interpretative flexibility expresses the fact that the same empirical data, produced or collected during a research, could be differently interpreted. Many factors, according to Collins, could explain such an 'interpretative flexibility'. One of them is tacit knowledge, about which Collins stresses particularly in his work. For him, the scientific practice is not resumed by formal rules or heuristics. Tacit knowledge can never be fully articulated or translated into a set of rules[3]. My reader might be surprised, but I

agree on the Stage 1 of EPOR. Indeed, it focuses on a fundamental element of science, as it runs in the laboratory: the contingency is irreducible in the day-to-day practices.

2.2. *The End of Debates between Competitive Interpretations of Experimental Data*

The Stage 2 of EPOR aims to offer an explanation about how the debates generated by the interpretative flexibility, introduced at the Stage 1, could be ended. In their empirical studies, the relativistic sociologists consider that the only factors which are relevant to conduct the debates to end are micro- and macro-social factors. Collins himself writes: “In each case (…), coherence and accumulation of experimental results are not enough to explain closure – reference to rhetoric, authority, institutional positions, and everything that comes under the « catchall » terms *interests* and *power*, is also required. (...) the consensual interpretation of day-to-day laboratory work is only possible within constraints coming from “*outside that work*’” (Collins 1982, 141-142).

If one accepts to consider this second stage of EPOR as relevant, one must agree with a particular vision of science concerning the role of intersubjectivity. Considering that intersubjectivity alone produces the content and the methods of science seems to be radical; but it is the conclusion when the constraints from the natural word are not considered as a decisive factor, whereas the mutual understanding of the subjects is essential. Collins (1981a, 3) is explicit about this point: “the approach ‘embraces an explicit relativism in which the natural world has a small or non-existent role in the construction of scientific knowledge”. And mutual understanding is characterized by him

as follows: “(…) mutual understanding seems to be possible even when nothing real is the subject matter. The quality of a poem or a picture, the number of angels that could dance on the head of a pin, or the cut of the emperor’s new clothes can all be discussed without there being any lumps in the world that correspond to them.” (1992, 174).

So should we renounce to an ideal of objectivity as the relativistic sociologists do? Should we restrict to nothing the role accorded to phenomena in the constitution of scientific knowledge? I don’t think so. We have to consider the fact that the researchers, in their laboratories, are competent enough to be reflexive about their practices. Indeed they are aware of the interpretative flexibility due to their idiosyncratic features, their tacit knowledge, the problems inherent to technical devices, and the impossibility to identically reproduce an experiment. Moreover, they are daily confronted with situations including uncertainty. To reduce this intrinsic limitation of the data’s collect, the scientists develop strategies to increase the degree of objectivity of their results. One of their strategies could be called “interinstrumentality”. Some authors proposed a quite similar point of view: Chang, 2001; Culp, 1995; Hacking 1983; Hudson, 1999; Nederbragt, 2003; Wimsatt, 1981. The fact to report to various different methods in order to confirm a hypothesis has several denominations: ‘robustness’ for Wimsatt (1981) or Chang (1995), ‘triangulation’ for Star (1986), ‘independence of route’ for Hudson (1999), ‘multiple derivability’ for Nederbragt (2003).

Three remarks in order to argue for the interest of my study. Firstly, it concerns the contemporaneous physics contrary to the previous which are about biology, except those of

Chang. Secondly, the biologists evoked in these studies try to locate the same entities. In surface science, the approach is a little bit different. Most frequently, the aim is not to locate the same objects on the images produced by different microscopes: it is impossible because the samples are deteriorated by the microscope's handling. For example, if a scientist uses a STM to observe cobalt atoms on a gold surface, the sample would be deteriorated. He will clean the gold substrate before evaporating cobalt. This new sample could be studied with a Transmission Electron Microscope (TEM) in order to confirm the presence of cobalt atoms. TEM allows to observe the atomic structure of a sample. In surface science, using various instruments aims to make sure that the properties expected from the studied objects are real. Thirdly, none of these studies points particularly that the interinstrumentality is a daily approach. Finally, few authors have considered the role of interinstrumentality during the phases of invention and diffusion of a new instrument contrary to the present study.

For me, 'interinstrumentality' could be characterized as a consecutive use of various instruments based on different physical principles in order to realize an experimental study. Each instrument provides a specific kind of information about the object studied. The information can be chemical, topographic, magnetic, electronic... The robustness of the interpretation is based for the scientists on the concordance between them[4].

3. Interinstrumentality and the Acceptance of the Scanning Tunneling Microscope

How could interinstrumentality play a role in the consensual agreement's process about the reliability of the STM? Designed in 1981 by Gerd Binnig and Heinrich Rohrer, the STM awarded the Nobel Prize of Physics in 1986. This microscope opens the door to what we call now nanotechnologies: single atoms on a surface can be displayed in three dimensions. The STM is based on the concept of quantum tunneling. When a conducting tip is brought very near to the surface to be examined, a voltage difference applied between the two can allow electrons to tunnel through the vacuum between them. This is the tunneling effect phenomenon in contradiction with the classical physics. The stakes were high: was it possible to build a macroscopic device which did not perturb the quantum effect observed only in the infinitesimal world? At that time, most of the scientists considered such ambition as unattainable.

Focusing on the process which conducts to the invention of STM, a surprising complexity emerges. Initially, Binnig and Rohrer did not expect to build a new microscope. They had insufficient knowledge about microscopy and surface science. At the request of their colleagues, whom studied insulating layers for electronic components, they researched a way to study finely the defaults of some materials. The original idea then was not to build a microscope but rather to perform spectroscopy locally on an area less than 100 Å in diameter. Only after many weeks, they realized that their probe could collect topographic information in addition to spectroscopic local information (Binnig & Rohrer 1986, 392). The STM was born.

In 1981, the first results were coldly received by the scientific community: in a private conversation, a researcher from the GSI told me: “These first images were literally incredible”. Some scientists went so far as to charge Binnig and Rohrer with fraud. “My personal opinion is, in that period, many scientists considered atom as a sacred object. It could never be handled. So an approach allowing to see and to handle atoms with such an accuracy was counterintuitive. The scientists did not want to consider such possibility. Indeed they refused to discuss it. It was like a taboo, like to talk about the devil.” (interview with Binnig and Rohrer, 2001)[5]. The first submission of the paper about the tunneling effect in open air was rejected. The referees argued that the study of the samples should be done under ultra-high vacuum in order to eliminate any possible contamination. And even if this procedure would be used, the samples could be corrupted before their insert in the ultra-high vacuum enclosure. A second critic of the referees was the insufficiency of the theory used to explain data. In response, Binnig and Rohrer argued in order to obtain the agreement of their colleagues.

The publications written since 1981 to 1985 reveal the different steps of this process. Besides scientific reports about their new data, each paper intended to respond to the detractors’ critics. My analysis reveals six strategies developed by Binnig and Rohrer:

(1) To consolidate the theoretical foundations of the phenomenon;

(2) To avoid all possible exogenous variables (vibrations’ reduction, vacuum’s improvement, better proceedings to prepare the samples…);

(3) To simplify the manipulation of the instrument, in order that other scientists could run the same experiment;

(4) To show the interest of the instrument in some other scientific areas;

(5) To study simple and well-known surfaces.

(6) To resort to interinstrumentality, in other words, to collect data with other scientific devices such as X-ray diffraction, TEM, etc.

These six strategies could be found in their papers during this period:

- In *Applied Physics Letters* (1982), the authors argue that the sources of vibration are enough controlled to interfere anymore with the collected data (Strategy 2)
- In *Physical Review Letters* (1982), in order to convince their peers that the data collected with the STM were not the product of their imagination, they use a common technique in Surface Science, the Low-energy electron diffraction (LEED) (Strategy 6);
- In *Surface Science* (1983), Binnig and Rohrer, who have technically simplified their instrument, try to convince the community of the surface science physicists of the STM's interest for their own researches. To do that, they study materials commonly used in surface science such as gold and silicon. To convince that their results are not a fancy, they compare them with others produced with instruments which have acquired a reliability in the community (Strategies 4, 5 and 6);

- In *Surface Science Letters* (1983), they use well-known materials and well-known instruments based on different physical principles such as TEM and X-rays diffraction (Strategies 5 and 6);
- In 1984, they are more confident in their method. Indeed, other researchers give a consistent theoretical basis to STM. From then on, Binnig and Rohrer endeavour to make their observations easier to reproduce. In their publication of 1984, in *Physica B*, they stress that the main source of problems is constituted by the metallic tip of the STM. They try to solve it (Strategies 1 and 3);
- Still in 1984, in *Surface Science*, they respond to a main critic: the risk of contamination during the sample's handling. They study gold's samples with different instruments (LEED and Auger spectroscopy of electrons - AES -), without any move of the samples. The same sample stays in the vacuum's enclosure and is examined successively with each instrument. Such handling reduces drastically the possibilities of contamination (Strategies 2, 5 and 6);
- In S*urface Science Letters* (1985), they increase the images of the silicon 7 x 7, and above all they corroborate their results appealing to AES and to LEED (Strategies 5 and 6);
- The final paper is published in *Scientific American*, so the STM can be known by a large public.

This brief historical background reveals the numerous trial-and-errors needed to pass from the first version of STM including a remarkable complexity to the easiest handling versions used today commonly in surface science laboratories. The numerous

improvements made to the first version of STM result partially from the objections expressed by the scientific community to Binnig and Rohrer: to put the STM under ultra-vacuum enclosure in order to restrict the contaminations, to integrate in their plan other instruments already used in 1980s as LEED, AES, etc. The debt contracted with the detractors is essential: they urge the two researchers to justify their theoretical principles and to modify their instrument. Binnig and Rohrer had to convince their colleagues providing good arguments in order to get the agreement of their peers. Each paper analyzed previously replied to a category of objections. The force of Binnig and Rohrer is not only in their attempt to test the reproducibility of their own experiments, but in their *quasi* systematic use of instruments commonly ran by their detractors in order to support their own argumentation. The scientists relied on well-known instruments used in surface science (as LEED, AES, TEM …) to confront their own results to those of others. This approach has two aims: to convince themselves of the reliability of their results and to convince their colleagues.

One could suppose that interinstrumentality is necessarily needed in such case, because of the controversial nature of the STM. But I want to underline the fact that interinstrumentality does not allow scientists to suppress all ambiguities and so to lead to absolute certainty, but it allows so “to decrease the risk of error” in their interpretation of the results. They increase the robustness of their conclusions.

When the process is achieved, has the STM a status which is once and for all justified? Could one use it as a black box? We answer both yes and no. The reliability of the results produced with a STM is no more discussed. A student could easily learn to run it.

However, when the aim is to use the instrument in certain conditions such as high resolution, for example, it is very difficult to master the various parameters and it needs a close knowledge of the instrument in order to obtain satisfactory results. In that case, the STM could not be considered as a black box.

In their day-to-day activities, the physicists are aware of the numerous uncertainties linked to the use of instruments as the STM, uncertainties which involve interpretative flexibility of data. The samples being irrevocably damaged with each observation, the experiments are strictly speaking not reproducible. The problem of reproducibility links in such a case to the mastery of concrete operations. Indeed, to master the technical factors (such as the variable quality of the vacuum, the reduction of the mechanic or electronic disturbances, etc.) is extremely problematic. Moreover, the tacit knowledge plays a great role in this kind of experiment (some manual know-how is needed in the polishing of a sample of metallic materials, in the preparation of the tip of a STM, contaminations, etc.). All variables have to be considered and to be mastered. An image of good quality needs frequently weeks to months to be produced with a STM. The surface science researchers are aware of the eminently problematic nature of the reproducibility so as the interpretative flexibility of the collected data. However, contrary to Collins, who argues that social factors are decisive for the end of the controversies, we can consider that such role is devolved in particular to interinstrumentality. Indeed the scientists are not confronted to endless controversies, because they can use strategies as interinstrumentality. In surface science, they use it *quasi*

systematically in order to reduce the interpretative flexibility, even if, in a first stage, the experiment seems to run perfectly.

4. Interinstrumentality and Day-to-Day Research in Surface Science

The case developed briefly in this part is based on ethnographical observations in a surface science laboratory. I consider such case as representative from the daily practices in the laboratory. The young scientist described here aims to get carbon nanotubes. To produce these, he elaborates a sample in which he believes that he has made grow such nanotubes. To study them, he uses a scanning electron microscope (SEM) which gives morphological information about the sample. The scientist sees, in his own terms, “a forest of spaghetti”. He considers that this “forest” corresponds to the expected nanotubes. At that point of his research, all seems to be coherent with his expectations and the handling occurred during the sample’s preparation. *A priori*, nothing constraints him to confirm his first SEM observations. But the scientist has learned that one sort of information produced with only one instrument could not be enough. So he has to try to corroborate his first observations. That is why he chooses to study his sample with a TEM in order to reveal the carbon atomic structure and to obtain chemical precisions at the same time. After many work sessions with TEM, the researcher does not have the expected result: he observes on the images things which seem to be nanotubes’ traces but the chemical analysis of these does not reveal carbon.

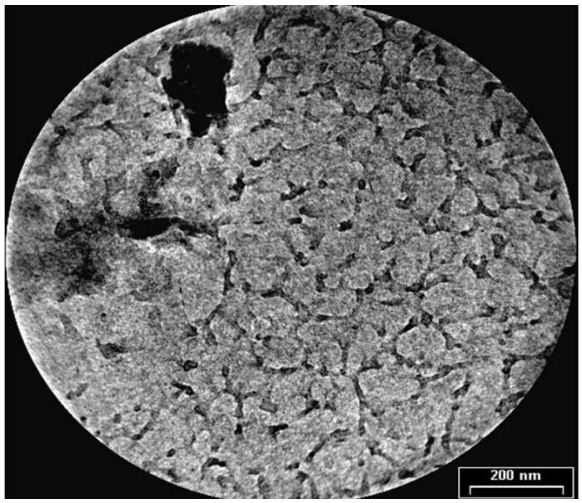


Figure 1: Nanotubes' traces or not? Micrograph performed with TEM. G. Erhet GSI .

How can this surprising fact be explained? Is it a problem of instrumental artifact? Or an artifact induced by the preparation of the sample? The two microscopes delivered conflicting informations. The physicist is now engaged in a laborious investigation in order to eliminate all the possible causes of artifacts. One has to precise that one major difficulty of such approach is the impossibility for the scientist to determine *a priori* if the incompatibility of the results between the two microscopes is due to a single artifact or by unfortunate conjunction of various artifacts. In order to succeed, one has to use an abductive reasoning (Peirce, 1903). This abductive stage is essential for two reasons: firstly, it is the creative side of a research's work and secondly, nothing guarantees that a researcher always produces fruitful abductive hypotheses. We observe that a problem could be generated by a conjunction of various causes. Moreover, our observation of the activities in this surface science laboratory leads us to assert that the competence to

formulate abductive hypotheses is variable from a researcher to another. This variation has direct effects on the quality of the researches, creating a degree of uncertainty.

Our scientist faced to a surprising phenomenon (the absence of nanotubes on the sample observed with TEM) should formulate hypotheses about its origin and corroborate them or not. In this case, hypotheses would be about the possible causes of artifacts: would the problem be linked to the preparation of the nanotubes? Would the microscopes or the computers used to produce data dysfunction? After a process of three months, the researcher finally finds that the preparation of nanotubes is the source of his problem. Contrary to his expectations, he did not produce carbon nanotubes, but another substance. In the case related here, the resort to interinstrumentality does not permit to confer to the first SEM images the status of proof. Consequently they could not be published and the study must run again including the constitution of new hypotheses.

Finally, we observe that an image produced by a microscope could never be considered by the scientists as a sufficient convincing proof. The scientists have to evaluate data with other instruments (TEM, STM, AES…) in order to produce other images and other results. Almost all of these instruments involve resorting to tacit knowledge and generate a part of interpretative flexibility. But what scientists are searching is the reliability of the information issued from the observations collected with the instruments. To obtain congruent information allows conferring a status of element of proof to such images instead of a status of simple and local coincidence. To resort to interinstrumentality is a

way to reduce the uncertainty. What the scientist aims to produce is a kind of Peircean cable constituted with many fibers (each fiber is, for example, an image). A cable is robust because it is made of many fibers and, unlike a chain, the solidity of which depends on its weakest link, the cable remains robust even if one or two of its fibers break. (Callebaut 1993, 57). This Peircean cable resorts to theoretical models, to intersubjectivity, and especially to interinstrumentality. When the Peircean cable is considered as enough consistent, *i.e* when the elements of proof are sufficiently matching and coherent among them to resist to objections, the scientists publish. However, a Peircean cable is always potentially subject to revision: new elements can be added, its composition can be criticized, it can be considered as insufficient or too partial. The texts of the referees are a particular and eloquent example of this last point. The referees stress frequently that the researchers assert such thesis relying on data insufficiently corroborated by other instruments.

5. Conclusion

With the two case studies briefly evocated above, I try to show that it is possible to enrich our understanding of the concept of objectivity, and most particularly when it concerns the results obtained in natural sciences. Traditionally, its definition refers mainly to intersubjectivity. I propose to complete it with a set of strategies, and especially the one I called interinstrumentality[6]. What the interinstrumentality questions are the conclusions proposed by Collins. For me, social factors are not essential to end the controversies due to

the interpretative flexibility. I could have two points of agreement with Collins: firstly, the choice of a strategy as interinstrumentality is based on a consensus among the scientists, and secondly the results produced with each instrument include social factors in a more or less large sense. But the convergence of results could not be reduced to social factors. For me, interinstrumentality increases the degree of reliability but cannot succeed to reach absolute reliability. It permits to conduct the investigation up to the point in which the hypotheses are considered as true “beyond all reasonable doubt”. So objectivity can no more be understood as a question of ‘all’ or ‘nothing’ but as a *continuum* according to Putnam (2003). Objectivity is a question of more or less.

References

Allamel-Raffin, Catherine., and Jean-Luc Gangloff. 2012. "Scientific Images and Robustness". In *Characterizing the Robustness of Science after the Practical Turn in Philosophy of Science*, eds. Léna Soler et al., 169-188. Frankfurt: Springer

Binnig, Gerd, Heinrich Rohrer, and *al.* 1982. "Tunneling through a Controllable Vacuum Gap." *Applied Physics Letters* 40: 178-180.

Binnig, Gerd, Heinrich Rohrer, and *al.*. 1982. "Surface Studies by Scanning Tunneling Microscopy." *Physical Review Letters*, 49: 57-61.

Binnig, Gerd, and Heinrich Rohrer. 1983. "Scanning Tunneling Microscopy." *Surface Science*, 126: 236- 244.

Binnig, Gerd, Heinrich Rohrer, and *al*. 1983. "(111)Facets as the Origin of the Reconstructed AU (110) Surfaces." *Surface Science Letters*, 131: L379-L384.

Binnig, Gerd, and Heinrich Rohrer. 1984. "Scanning Tunneling Microscopy." *Physica B*, 127: 37-45.

Binnig, Gerd, Heinrich Rohrer, and *al*. 1984. "Real-Space Observation of the Reconstruction of AU (100)." *Surface Science*, 144: 321-335.

Binnig, Gerd, Heinrich Rohrer, and *al*. 1985. "Revisiting the 7 x 7 Reconstruction of Si (111)." *Surface Science Letters*, 157: L373-L378.

Binnig, Gerd, and Heinrich Rohrer. 1985. "The Scanning Tunneling Microscope." *Scientific American*, 253: 40-46.

Binnig, Gerd., and Rohrer, Heinrich. 1986. "Nobel Lecture" http://www.nobel.se/physics/laureates/1986/binnig-lecture.pdf.

Callebaut, Werner. 1993. *Taking the Naturalistic Turn or How Real Philosophy of Science is Done*. Chicago: University of Chicago Press.

Chang, Hasok. 1995. "Circularity and Reliability in Measurement." *Perspectives on Science*, 3 (2): 153-172.

Chang, Hasok. 2001. "Spirit Air, and the Quicksilver: The Search for the Real Temperature." *Historical Studies in the Physical and Biological Sciences*, 3, 260-284.

Collins, Harry. M. 1981a. "Stages in the Empirical Programme of Relativism." *Social Studies of Science,* 11 (1): 3-10.

Collins, Harry. M. 1981b. "Son of Seven Sexes: The Social Destruction of a Physical Phenomenon." *Social Studies of Science*, 11 (1): 33-62.

Collins, Harry. M. 1982. "Special Relativism: The Natural Attitude." *Social Studies of Science*, 12 (1): 139-143.

Collins, Harry. M. 1985. "An Empirical Relativist Programme in the Sociology of Scientific Knowledge." In *Science Observed. Perspectives on the Social Studies of Science*, ed. K. Knorr-Cetina, and M. Mulkay, 85-113. London & New Delhi & Beverly Hills: Sage.

Collins, Harry. M. 1992. *Changing Order. Replication and Induction in Scientific Practice*. London: Sage Publication.

Collins, Harry. M. 2004. *Gravity's Shadow. The Search for Gravitational Waves*. Chicago: University of Chicago Press.

Collins, Harry. M. 2010. *Tacit and Explicit Knowledge*, Chicago: University of Chicago Press.

Collins, Harry. M., and Trevor Pinch. 1993. *The Golem.* Cambridge: Cambridge University Press.

Culp, Sylvia. 1995. "Objectivity in Experimental Inquiry: Breaking Data-technique Circles." *Philosophy of Science*, 62:430-450.

Daston, Lorraine. 1992. "Objectivity and the Escape from Perspective." *Social Studies of Sciences*, 22: 597-618.

Franklin, Allan. 2009. "Experiment in Physics, Stanford Encyclopedia of Philosophy." http://plato.stanford.edu/entries/physics-experiment/.

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

Hudson, Robert G. 1999. "Mesosomes: A Study in the Nature of Experimental Reasoning." *Philosophy of Science*, 66: 289-309.

Megill, Allan. 1994. "Introduction." In *Four Senses of Objectivity, Rethinking Objectivity*, by Allan Megill, 1-20. London: Duke University Press.

Nederbragt, Hubertus. 2003. "Strategies to Improve the Reliability of a Theory: the Experiment of Bacterial Invasion into Cultured Epithelial Cells." *Studies in History and Philosophy of Biological and Biomedical Sciences*, 34: 593-614.

Peirce, C. S. (1903), *Harvard lectures on pragmatism*, *Collected Papers* v. 5,§ 188-189.

Putnam, Hilary. 2003. “Pragmatisme et connaissance scientifique." In *Cent ans de philosophie américaine*, ed. Jean-Pierre Cometti and Claudine Tiercelin, 135-155. Pau : Presses Universitaires de Pau.

Toumey, Chris. 2011. “Compare and Contrast as Microscopes Get Up Close and Personal.", *Nature Nanotechnology*, 6: 191-193.

Star, Susan L. 1986. "Triangulating Clinical and Basic Research: British Localizationists, 1870-1906." *History of Science*, 29: 29-48.

Soler, Léna, Emiliano Trizio, Thomas Nickles, and William C. Wimsatt. 2012. *Characterizing the Robustness of Science after the Practical Turn*, Frankfurt:Springer.

Wimsatt, William C. 1981. "Robustness, Reliability and Overdetermination." In *Scientific Inquiry and the Social Sciences*, ed. M.B. Brewer, and B.E. Collins, 124-163. San Francisco: Jossey-Bast Publishers.

http://hrst.mit.edu/hrs/materials/public/Binnig&Rohrer.htm, site of Dibner Institute for History of Science and Technology.

When to expect violations of causal faithfulness and why it matters

Holly Andersen

Simon Fraser University

holly_andersen@sfu.ca

Abstract: I present three reasons why philosophers of science should be more concerned about violations of causal faithfulness (CF). In complex evolved systems, mechanisms for maintaining various equilibrium states are highly likely to violate CF. Even when such systems do not precisely violate CF, they may nevertheless generate precisely the same problems for inferring causal structure from probabilistic relationships in data as do genuine CF-violations. Thus, potential CF-violations are particularly germane to experimental science when we rely on probabilistic information to uncover the DAG, rather than already knowing the DAG from which we could predict the right experiments to 'catch out' the hidden causal relationships.

Wordcount, including references, abstract, and footnotes: 4973

1. Introduction

Several conditions must be met in order to apply contemporary causal modeling techniques to extract information about causal structure from probabilistic relationships in data. While there are slightly different ways of formalizing these requirements, three of the most important ones are the causal Markov, causal modularity, and causal faithfulness conditions. Potential failures of the first two of these conditions have already been the subject of discussion in philosophy of science (Cartwright 1999, 2002, 2006; Hausman and Woodward 1999, 2004; Steel 2006; Mitchell 2008; Woodward 2003, 2010). I will address failures in the third condition, causal faithfulness, and argue that failures of this condition are likely to occur in certain kinds of systems, especially those studied in biology, and are the most likely to cause trouble in experimental settings.

Faithfulness is the assumption that there are no precisely counterbalanced causal relationships in the system that would result in a probabilistic independence between two variables that are actually causally connected. While faithfulness failures have been discussed primarily in the formal epistemology literature, I will argue that violations of faithfulness can impact experimental techniques, inferential license, and issues concerning scientific practice that are not exhausted by the formal epistemology literature.

In particular, a formal methodological perspective might suggest a distinction between genuine and merely apparent failures of CF, such that supposed examples of CF-violating systems are not 'really' CF-violating, but merely close. But as I will argue, this

distinction is not epistemically justifiable in experimental settings: we cannot distinguish between genuine and merely apparent CF violations unless we already know the underlying causal structure; without this information, merely apparent and genuine CF violations will be indistinguishable. Violations of CF faithfulness are particularly germane to experimental science, since CF is the assumption that takes us from probabilistic relationships among variables in the data to the underlying causal structure. In contrast, for instance, the Causal Markov condition takes us from causal structure to predicted probabilistic relationships. Going from data to underlying causal structure is the most common direction of inference from the epistemic vantage point of science. Rather than beginning by knowing the true causal graph of the system in question to predict probability distributions, experiment moves from probabilistic relationships to the underlying causal structure.

This means that failures of CF arguably have the most potential for wreaking havoc in experimental settings, and have interesting methodological consequences for the practice of science: we should expect to find epistemic practices that compensate for CF-violations in fields that study systems where faithfulness is likely to fail. Thus, these conditions are of interest not only to those working on formal modeling techniques, but also to broader discussions in philosophy of science, especially those that concern epistemic practices in the biological, cognitive, or medical sciences.

2. Violations of the Causal Faithfulness Condition

Violation of CF occurs when a system involves precisely counterbalanced causal relationships. These causal relationships appear "invisible" when information about

conditional and unconditional probabilities is used to ascertain a set of possible causal directed acyclic graphs (DAGs) that are consistent with data from that system. More precisely:

> Let *G* be a causal graph and *P* a probability distribution generated by *G*. <*G, P*> satisfies the Faithfulness Condition if and only if every conditional independence relation true in *P* is entailed by the Causal Markov Condition applied to *G*. (Spirtes, Glymour, and Scheines 2000, 31)

> One can think of faithfulness as the converse of the Causal Markov condition: faithfulness says that given a graph and associated probability distribution, the only independence relations are those that follow from the Causal Markov condition alone and not from special parameter values… (Woodward 2003, 65)

Informally, variables should only be probabilistically independent if they are causally independent in the true causal graph; when causal relationships cancel each other out by having precisely counterbalanced parameter values, the variables are probabilistically independent, but not causally independent. Thus, in systems that have CF-violating causal relationships, the probabilistic relationships between variables include independencies that do not reflect the actual causal relationships between those variables.

Probabilistic relationships are used to generate possible causal graphs for the system. There may be multiple distinct causal graphs which all imply the observed set of

probabilistic relationships. The candidate graphs can then be used to generate further interventions in the system that will distinguish between the graphs; if two candidate graphs make different predictions for the consequences of an intervention on variable A, then performing this intervention on A should return an answer as to which of the candidates graphs matches the observed results. The use of probabilistic data to generate candidate causal graphs that can then be used to suggest further interventions can save huge amounts of time and energy by focusing on a few likely candidates from an indefinitely large number of candidate causal structures.

DAGs of causal faithfulness violations may take several forms. For example:

Figure 1a

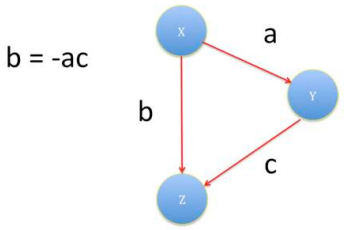
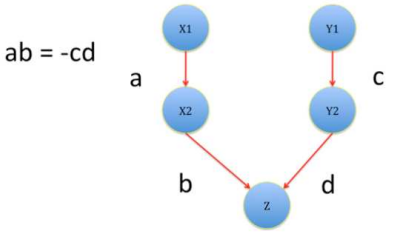


Figure 1b



Some authors (Pearl 2000, Woodward 2010) rely on a stronger constraint, causal stability, which requires that probabilistic independence relationships be stable under perturbation of parameter values across some range, to eliminate "pathological" (i.e. CF-violating) parameter values.

Definition 2.4.1 Stability:

> Let I(P) denote the set of all conditional independence relationships embodies in P. A causal model M = <D, Θ> generates a stable distribution if and only if P(<D, Θ>) contains no extraneous independences – that is, if and only if I(P(<D, Θ>)) ⊆ I(P(<D, Θ`>)) for any set of parameters Θ`. (Pearl 2000)

Violating causal stability would require a system to respond to changes in one parameter value with compensating changes in another parameter, so that the values remain exactly counterbalanced for some range of values.

The potential for CF-violations to reduce the reliability of methods for extracting causal structure from data is well-known in formal epistemology. However, I will argue that philosophers of science in general should pay more attention to such violations; understanding the difficulties that CF-violations pose will enhance our ability to accurately characterize features of experimental practice, and should be included in normative considerations regarding evidence and inference. The main arguments in this paper can be summarized in three brief points:

> (1) Even if CF-violating systems are measure 0 with respect to the set of causal systems with randomly distributed parameter values, this does not imply that we will only encounter them with vanishing probability. CF-violating systems may be of particular interest for modeling purposes compared to non-CF-violating systems, in particular because certain kinds of systems may have structural features that render CF-violating parameter values more likely.

(2) As an example of point 1, structural considerations regarding dynamically stable systems that are the result of evolutionary processes should lead us to expect CF-violations in various biological systems. For systems that have evolved to maintain stable equilibrium states against external perturbation, we should also expect violations of the stronger condition, causal stability. I briefly present an example of this: mechanisms for salinity resistance in estuary nudibranchs.

(3) 'Apparent' CF-violations in equilibrium-maintaining systems can be generated in certain experimental conditions even though the actual causal relationships in question may not be exactly balanced. Some measurement circumstances will result in a data set that violates CF, even if the actual system being measured does not genuinely violate CF. We should be as concerned with merely apparent as with genuine CF-violations, since both kinds of violations lead to the same difficulties for moving from probabilistic relationships in data to accurate DAGs of systems.

These three points highlight why philosophers of science in general should be concerned: causal systems may not genuinely violate CF, but yet pose the same problems for experimental investigations as if they did. Apparent CF-violations occur when systems do not in principle violate CF but appear to due to measurement issues connected with data-gathering. In both genuine and merely apparent CF-violations, probabilistic relationships in the data will suggest a set of candidate causal graphs that are inaccurate; as a result, further interventions will yield conflicting answers. Scientists could in principle 'catch out' these merely apparent CF-violations *if they knew exactly how to test for them*. But to

do this, they would need the DAG, and this is the information that they lack when proceeding from the data to underlying causal structure. When we have incomplete knowledge of the causal structure of the system under investigation, we lack this ability to distinguish between merely apparent and genuine CF-violations. Both raise the same problems.

3. The measure of CF-violating systems

Spirtes, Glymour, and Scheines (2000) offer a proof that CF-violating systems are Lebesgue measure 0 with respect to possible causal systems, while non-CF-violating systems are measure 1. "The parameter values—values of the linear coefficients and exogenous variances of a structure—form a real space, and the set of points in this space that create vanishing partial correlations not implied by the Markov condition have Lebesgue measure 0" (41). From this, they conclude that we are vanishingly unlikely to encounter CF-violating systems, and so proceed on the initial presumption that any given causal system is not CF-violating. This proof may be part of the reason why comparatively little attention has been paid to causal faithfulness compared to the causal Markov and modularity conditions. However, the fact that CF-violating systems are measure 0 in this class does not imply that we will not encounter them with any frequency.

To motivate this, consider an analogy with rational numbers. They are also measure 0 with respect to the real numbers, while irrational numbers are measure 1. And, there are circumstances under which we are vanishingly unlikely to find them. If a random real number were to be chosen from the number line, the probability that we will

draw an irrational number is so overwhelming as to warrant ignoring the presence of rational numbers. However, this does not imply that rational numbers are unlikely to be encountered *simpliciter*: bluntly put, we don't 'encounter' the numbers by randomly drawing them from the number line. Rational numbers are encountered overwhelmingly more often than one would expect from considering only the proof that they are measure 0 with respect to real numbers.

The Spirtes, Glymour, and Scheines proof assumes that all parameter values within the range of a continuous variable are equally probable (Zhang and Spirtes 2008). Without this assumption, one can't presume that the CF-violating values are vanishingly unlikely. For instance, this assumption does not hold for systems that involve equilibrium-maintaining causal mechanisms. Such mechanisms work to maintain counterbalanced parameter values, rendering it much more likely that parameter values will result in CF-violations.

It is true that if causal systems took on parameter values randomly from their range, we would expect to encounter CF-violating systems with vanishingly small probability, and in that scenario, we could safely ignore CF-violations as a real possibility on any given occasion. However, some systems survive, and become scientifically interesting targets for investigation, precisely because they achieve long-term dynamic equilibrium using mechanisms that rely on balanced parameter values. In such systems, the parameter values are most certainly not indifferently probable over their range. In fields like biology, neuroscience, medicine, etc., we are disproportionately interested in modeling systems that involve equilibrium maintaining mechanisms. This suggests that our modeling interests are focused on CF-violating systems in a way that is

disproportionate to their measure when considered against all possible causal systems. Thus, we cannot conclude from the fact that CF-violating parameter values have measure 0 with respect to all possible parameter values that we will not encounter such violations on a regular basis.

Zhang and Spirtes (2008) discuss some circumstances in which systems may violate CF. However, their discussion makes it seem like CF-violations occur primarily in artificial or constructed circumstances. One such example is homeostatic systems, which maintain equilibrium against some range of perturbations, such as thermostats maintaining a constant temperature in a room. Zhang and Spirtes demonstrate that CF can be replaced with two distinct subconditions, that, taken together, provide almost the same inferential power as causal faithfulness. If systems violate only one of these subconditions, such violations can be empirically detected. This is an extremely useful result, and increases the power of Bayes' nets modeling to recover DAGs from data. However, this result should not be taken as resolving the problem.

In particular, their use of a thermostat as example of a homeostatic system does not do justice to the incredibly complex mechanisms for homeostasis that can be found in various biological systems. Considering these more sophisticated examples provides a clearer view of the potential problems involved in modeling such systems under the assumption of causal faithfulness.

4. Evolved dynamical systems and equilibrium-maintaining mechanisms

The tendency for evolved systems like populations, individual organisms, ecosystems, and the brain to involve precisely balanced causal relationships can be easily

explained by the role these balanced relationships play in maintaining various equilibrium states (see, for instance, Mitchell 2003, 2008). Furthermore, the mechanisms by which organisms maintain internal equilibrium with respect to a huge variety of states will need to be flexible. They need to not simply maintain a static equilibrium, but respond to perturbation from the outside by maintaining that equilibrium. This means that many mechanisms for equilibrium maintenance will have evolved to keep an internal state fixed over some range of values in other variables, not merely for a single precise set of values. Any system that survives because of its capacity to maintain stability in the face of changing causal parameters or variable values will be disproportionately likely to display CF-violating causal relationships, and, more strongly also violate causal stability.

An intriguing example is nudibranchs, commonly known as sea slugs (see especially Berger and Kharazova 1997). Many nudibranchs live in ecosystems such as reefs, where salinity levels in the water change very little. These nudibranchs are stenohaline: able to survive within a narrow range of salinity changes only. In cases where salinity levels vary over narrow ranges, nudibranchs respond to changes in salinity levels by a cellular mechanism for osmoregulation, where cells excrete sodium ions or take in water through changes in cell ion content and volume. This mechanism provides tolerance, but not resistance, to salinity changes, because it maintains equilibrium by exchanging ions and water with the surrounding environment. In cases of extremely high or low salinity, this mechanism will cause the animal to extrude too much or take in too much (this is why terrestrial slugs die when sprinkled with salt).

Euryhaline nudibranchs, found in estuary environments where saline levels may vary dramatically between tides and over the course of a season or year, display a much

higher level of resistance to salinity changes. There is a pay-off, in the form of increased food sources with reduced competition for nudibranchs that are able to withstand the changing saline levels. But in these environments, the osmoregulatory mechanism for salinity tolerance is insufficient. A further mechanism has evolved in nudibranchs (and in molluscs more generally) for salinity resistance in conditions of extreme salinity variations in the external environment. These two mechanisms for salinity regulation in euryhaline nudibranchs are fairly independent. The osmoregulation mechanism is supplemented with an additional mechanism which involves hermeticization of the mantle, which prevents water and ion exchange with the outside environment.. This can accommodate changes in salinity that take place over fairly short periods of time, since salinity levels can change dramatically over the course of an hour. Instead of maintaining blood salinity at the same level as the outside environment, this additional mechanism allows the organism to maintain an internal salinity level that differs from that of its environment. Mantle hermeticization and osmoregulation are distinct mechanisms, but in contexts of extremely high or low salinity, they will both act such that the variables of external and internal salinity are independent

Further, there are two distinct mechanisms in muscle cells that work in coordination in extreme salinity cases to maintain a balance of ions inside the muscle cell. The concentration of these ions, especially sodium and potassium, can change dramatically in low or high salinity levels. There are two ion pumps in the cell that maintain overall ion concentration at equilibrium across a fairly substantial range of salinity variation in the external environment. Even though external salinity has several causal effects on the internal ion balance of a cell, these two variables will be probabilistically independent for

a range of external salinity values (in particular, for the range in which the organisms are naturally found).

> The ion balance of muscle cells during adaptation to various salinities could not be achieved by virtue of the Na/K-pump alone, removing sodium and accumulating potassium. As it is clear from the data obtained, the concentration of both ions drops at low salinity and increases at high salinity. Therefore, the effective ion regulation in molluscan cells can be provided only by cooperative action of two pumps – the Na/K-pump and Na,Cl-pump, independent of potassium transport. (Berger and Karazova 1997, 123-4)

There are several points that this example illustrates. The first is that of the comparative probability that a complex system, such as an organism like a nudibranch, will display CF-violating causal relationships in the form of mechanisms that maintain equilibrium. Consider the (Spirtes, Glymour, and Scheines 2000) proof that assumes that all parameter values are equally likely. We can see how this falls apart in the case of evolved systems. Let's grant that, in some imaginary past history, all the parameter values for mechanisms such as these two ion pumps were equally likely. This would have resulted in a vast number of organisms that ended up very rapidly with internal ion imbalances and then (probably rather immediately) died. The organisms that managed to stick around long enough to leave offspring were, disproportionately, those with mechanisms that were precisely counterbalanced to maintain this internal equilibrium. Having CF-violating mechanisms would be a distinct advantage. The same applies for

other important equilibrium states –organisms with less closely matched values are less capable of maintaining that equilibrium state. Insofar as these are important states to maintain, it becomes extremely probable that. Over time, those with the closest matches for parameter values will be more likely to survive. Thus, even if we grant the assumption (already unlikely in this context) that all parameter values start out as equally likely, we can see how rapidly the CF-violating ones would come to be vastly overrepresented in the population.

The second point it illustrates is how such sophisticated equilibrium-maintaining mechanisms can violate CF in a much more problematic way than the comparatively simplistic thermostat example considered by Zhang and Spirtes.[1] Finally, note that the two ion pump mechanisms are not balanced merely for a single external salinity value: they are balanced for a range of values. Thus, this example violates not merely CF but also the stronger condition of causal stability.[2]

I am certainly not claiming that all causal relationships in such systems will violate CF or causal stability. But it is possible that, for any given system that involves equilibrium-maintaining mechanisms, and especially for those with sophisticated evolved equilibrium-maintaining mechanisms, there will be at least some causal relationships in

[1] Note that a DAG representing the two mechanisms for the ion pumps, connecting external salinity levels as a variable to a variable representing internal ion balance in muscle cells, is not of the triangular form that is potentially detectable using the methods in Zhang and Spirtes (2008).

[2] This example also provides weight to the Russo-Williamson thesis, that information about probabilistic relationships requires supplementation with information about underlying mechanisms in order to justify causal claims. These examples suggest how investigation into mechanisms for equilibrium-maintenance compensate for the methodological issues that CF violations generate; we would expect the Russo-Williamson thesis to hold particularly of systems liable to violate CF.

the system that violate either or both of these conditions. This changes the stance we take at the beginning of an investigation: rather than starting from the assumption that CF-violations are vanishingly unlikely, and only revisiting this assumption in the face of difficulties, we should start investigations of such systems with the assumption that it is highly likely that there will be at least one such spurious probabilistic independence.

5. Apparent CF-violations and their experimental consequences

Consider a possible response to the argument in the previous section. One might be concerned that the examples I offer do not involve genuine CF-violations–when examined more closely, it may turn out that the causal relationships in questions are not exactly balanced, but merely close. This response might involve the claim that even in the case of biological systems, CF is not genuinely violated, because there are slight differences in parameter values that could be identified, especially if one performed the right interventions on the systems to 'catch out' the slight mismatch in parameter values. Or, by taking recourse to causal stability, one might say that while the equilibrium state of some systems involves precisely counterbalanced causal relationships, in the case of perturbation to that equilibrium, these relationships will be revealed. Perturbation of systems that return to equilibrium would thus be a strategy for eliminating many (or most) merely apparent CF-violations.

Answering this challenge brings us to the heart of why CF-violations deserve broader discussion. Considered from a formal perspective, there is a deep and important difference between systems that actually violate CF, or causal stability, and those that do not. This fact motivates a response to merely apparent CF-violations that takes them to be

not methodologically problematic in the same way that genuine ones are. But the ways in which merely apparent CF-violations can be 'caught out' generally will require information about the DAG for the system, in order to predict precisely which variables should be intervened on, within what parameter ranges, in order to uncover closely-but-not-exactly matched parameter values. While it is in principle possible to do this, it requires knowing precisely which intervention to perform, and it is this information that will be lacking in a large number of experimental situations where we don't already have the DAG for the system, since that is what we are trying to find.

Thus, a particular data set drawn from a target system for which investigators are seeking the DAG may have spurious conditional independencies between variables (i.e. violate CF) even though in the true DAG, those parameters are not precisely balanced. In other words, depending on how the data is obtained from the system, the data set may violate CF even though the system itself doesn't. How could this happen? There are a soberingly large number of ways in which a data set can be generated such that a merely apparent CF-violation occurs. The point to note here is that *merely apparent violations will cause exactly the same problems for researchers as would genuine CF-violations*. There are methodological issues in dynamically complex systems such that a non-CF-violating system may nevertheless result in a dataset that is CF-violating. Here are some ways in which this may happen.

The first is quite obvious: parameter values that are not exactly opposite may nevertheless be close enough that their true values differ by less than the margin of error on the measurements. Consider the parameter values in diagram 1a. A genuine CF-violation will occur if $a=-bc$. However, an apparent CF-violation will occur if $a\pm\varepsilon_1=-$

$bc \pm \varepsilon_2$. Concerns about the precision of measurements and error ranges are well-known, but it is useful to consider them here with respect to the issue of causal faithfulness as another way to flesh out their role in investigatory practices.

Two other ways in which apparent CF-violations may occur concern temporal factors which may play a key role in the 'catching' of equilibrium-balanced causal relationships. Temporal factors can distinguish systems with or without causal stability, for instance, a CF-violating system that is fragilely balanced.

Consider the time scale of a system that involves balanced causal relationships for the purposes of restoring and maintaining some equilibrium state: this may be on the order of milliseconds for some cellular processes, tens to hundreds of milliseconds for many neurological processes, minutes to days for individual organisms. After a perturbation takes place, the system will re-establish equilibrium during that range of time. In order to successfully 'catch' the counterbalanced causal relationships in the act of re-equilibrating, the time scale of the measurements must be on a similar or shorter time scale. If the time scale of measurements is long with respect to the time scale for re-establishing equilibrium, these balanced causal relationships will not be caught.

This basic point about taking state change data from dynamic processes has particular implications for CF-violations. For processes that re-equilibrate after 50 ms, for instance, a measurement device that samples the process at higher time scales, such as 500ms, will miss the re-equilibration. Thus, even though the system does not violate causal stability, it will behave as if it does, as it will appear that there is a conditional independence between two variables across some range of values, namely, the range between the initial state and the state to which the system was perturbed. In particular, if

we do not know what the time scale is, or is likely to be, for re-equilibration, we cannot ensure that a persisting probabilistic independence between two variables in question is genuine or a consequence of an overly fast re-equilibration timescale.

Not only does comparative time scales matter for apparent CF-violations; there are also possibilities for phase-matched cycles that that will make a non-CF-violating oscillating system appear to violate CF. Some systems develop equilibrium mechanisms that result in slight oscillations above and below a target state. If the measurements from this system are taken with a frequency that closely matches that of the rate of oscillation, then the measurements will pick out the same positions in the cycle, essentially rendering the oscillation invisible. This would constitute an apparent CF-violation as well.

Predicting possible CF-violations, real or apparent, requires information about the dynamic and evolved complexity of the systems in question, the particular equilibrium states they display, the time scale for re-establishment of equilibrium compared with the time scale of measurement, and/or the cycle length for cyclical processes.

6. Conclusion

To summarize briefly: some kinds of systems, especially those studied in the so-called 'special sciences', are likely to display the kinds of structural features that lead to CF-violations, such as mechanisms for equilibrium maintenance across a range of variable values. Some systems that do not have CF-violating DAGs may nevertheless generate CF-violating data sets. When we are considering the inferences made from probabilistic relationships in data to a DAG for the underlying system, and do not already have the DAG in hand, we cannot distinguish between genuine and merely apparent CF-

violations; both will cause the same epistemic difficulties for scientists, which is why merely apparent CF-violations deserve broader attention.

It's important to note that I am not discounting the extraordinary achievements in formal epistemology and causal modeling that have marked the last two decades of research on this topic. The steps forward in this field have been monumental, including the development of methods by which to reduce some of the issues arising from CF-violations (such as Zhang and Spirtes 2008). Rather, my goal is to clarify the ways in which apparent CF-violations can arise, the kinds of structural features a system might display that would increase the likelihood of CF-violation, and to bring this issue from discussion in formal epistemology into consideration of scientific practice more broadly.

References

Berger, V.J., and A.D. Kharazova. 1997. "Mechanisms of Salinity Adaptations in Marine Mollusks." *Hydrobiologia* 355 (1-3): 115-126.

Cartwright, Nancy. 1999. "Causal Diversity and the Markov Condition." *Synthese* 121 (1-2): 3-27.

Cartwright, Nancy. 2002. "Against Modularity, the Causal Markov Condition, and Any Link between the Two: Comments on Hausman and Woodward." *The British Journal for the Philosophy of Science* 53 (3): 411-453.

Cartwright, Nancy. 2006. "From Metaphysics to Method: Comments on Manipulability and the Causal Markov Condition." *The British Journal for the Philosophy of Science* 57(1): 197-218.

Hausman, Daniel M. and James Woodward. 1999. "Independence, Invariance and the Causal Markov Condition." *The British Journal for the Philosophy of Science* 50 (4): 521-583.

Hausman, Daniel M. and James Woodward. 2004. "Modularity and the Causal Markov Condition: A Restatement." *The British Journal for the Philosophy of Science* 55 (1): 147-161.

Mitchell, Sandra D. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge Studies in Philosophy and Biology: Cambridge University Press.

Mitchell, Sandra D. 2008. "Exporting Causal Knowledge in Evolutionary and Developmental Biology." *Philosophy of Science* 75 (5): 697-706.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Russo, Federica and Jon Williamson. 2007. "Interpreting Causality in the Health Sciences." *International Studies in the Philosophy of Science* 21 (2): 157-170.

Steel, Daniel. 2006. "Indeterminism and the Causal Markov Condition." *The British Journal for the Philosophy of Science* 56 (1): 3-26.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Cambridge, MA: The MIT Press.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, James. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology and Philosophy*. 25 (3): 287-318.

Zhang, Jiji and Peter Spirtes. 2008. "Detection of Unfaithfulness and Robust Causal Inference." *Minds and Machines* 18 (2): 239-271.

Emergence in Effective Field Theories

Jonathan Bain
Polytechnic Institute of New York University
6 Metrotech Center, Brooklyn, NY 11201

Abstract. This essay considers the extent to which a concept of emergence can be associated with Effective Field Theories (EFTs). I suggest that such a concept can be characterized by microphysicalism and novelty underwritten by the elimination of degrees of freedom from a high-energy theory, and argue that this makes emergence in EFTs distinct from other concepts of emergence in physics that have appeared in the recent philosophical literature.

1. Introduction

An effective field theory (EFT) of a physical system is a description of the system at energies low, or distances large, compared to a given cutoff. EFTs are constructed *via* a process in which degrees of freedom are eliminated from a high-energy/short-distance theory. Formulating a concept of emergence for EFTs is important for at least two reasons. First, EFTs play essential roles in contemporary physics: many authors believe the Standard Model of particle physics is an EFT, and most if not all condensed matter systems can be described by EFTs. Second, the types of physical systems that can be described by EFTs have been associated with various concepts of emergence in the recent philosophical literature: Mainwood (2006) suggests that the "new emergentism" of condensed matter physicists (*e.g.*, Anderson 1972, Laughlin and Pine 2000) can be characterized by microphysicalism and novelty underwritten by the physical mechanisms of spontaneous symmetry breaking and universality. Morrison (2012) similarly stresses the role of spontaneous symmetry breaking as essential to a concept of emergence, while Batterman (2011) focuses on universality. On the other hand, Wilson (2010) claims an appropriate concept of emergence should be based on the elimination of degrees of freedom from a theory in physics. I will suggest that while a concept of emergence appropriate for EFTs shares aspects of these views, it is distinct from them.

The plan of the essay is as follows. Section 2 reviews the steps involved in the construction of an EFT, section 3 offers an interpretation of EFTs from which section 4 extracts a concept of emergence based on the notions of microphysicalism and novelty. Finally, section 5 compares this concept with recent discussion of emergence in the philosophical literature.

2. EFTs and the Elimination of Degrees of Freedom

The concept of emergence I wish to associate with EFTs will ultimately be based on the elimination of degrees of freedom from a field theory in physics. I will take a degree of freedom associated with a theory to be a parameter that needs to be assigned a value in order to provide a dynamical state description of a physical system described by the theory. A dynamical state

description is a description of the system at an instant in time that, in conjunction with an equation of motion, determines a future or a past state. Thus, for example, a dynamical state description of a free classical particle governed by a second-order partial differential equation of motion (Newton's second law, for instance) is specified by the values of its position and momentum. In three spatial dimensions, this amounts to 6 degrees of freedom. A dynamical state description of a free classical field $\phi(x)$ governed by a second-order partial differential equation of motion is specified by the values that $\phi(x)$ and its first derivative $\partial_\mu\phi(x)$ take at every point x of spacetime, which amounts to an infinite number of degrees of freedom.

For some field theories, degrees of freedom associated with high energies (or short distances) can be eliminated in such a way that the result is an effective field theory that produces the same predictions as the original when restricted to low energies (large distances). One advantage of using the effective theory is that it makes calculations more tractable. Moreover, many quantum field theories can only be solved *via* perturbative expansions which contain divergent integrals at high energies. For these theories, the construction of a low-energy effective theory provides not just a practical way of avoiding these divergences, but a conceptual framework on which to build an interpretation of what these theories are telling us about the world. This construction proceeds in two steps:

(I) The high-energy degrees of freedom are identified and integrated out of the Lagrangian density representing the theory.

This first step assumes that the theory is encoded in a Lagrangian density $\mathcal{L}[\phi]$, which is a functional of a field variable $\phi(x)$.[1] This means that $\mathcal{L}[\phi]$ depends on all the possible functional forms the field can take, each form $\phi(x)$ taking values at all spacetime points x. Each such form of $\phi(x)$ represents a possible field configuration of field values; *i.e.*, a possible way the field could be spread over spacetime. To identify the high-energy degrees of freedom, one first choses an appropriate energy cutoff Λ and then decomposes the field variable into high- and low-energy parts, $\phi(x) = \phi_H(x) + \phi_L(x)$, where $\phi_H(x)$ and $\phi_L(x)$ are associated with momenta greater than and less than Λ, respectively. Once this is done, the high-energy degrees of freedom $\phi_H(x)$ are integrated out of the generating functional Z constructed from $\mathcal{L}[\phi_H, \phi_L]$,

$$Z = \int \mathcal{D}\phi_L \, \mathcal{D}\phi_H \, e^{i\int d^4x \, \mathcal{L}[\phi_L, \phi_H]} = \int \mathcal{D}\phi_L \, e^{i\int d^4x \, \mathcal{L}_{eff}[\phi_L]} . \qquad (1)$$

This functional integral is taken over all possible field configurations of the high-energy degrees of freedom $\phi_H(x)$. This literally eliminates these degrees of freedom from the Lagrangian density by replacing them with appropriate configurations of the remaining degrees of freedom (conceptually, in the same way a variable y is eliminated from an algebraic equation $ax + by = c$, by replacing it with an appropriate relation for the remaining variable: $y = (c - ax)/b$). The result of this is an effective Lagrangian density $\mathcal{L}_{eff}[\phi_L]$ that depends only on the low-energy degrees of freedom $\phi_L(x)$.

[1] In general a Lagrangian density of a field theory $\mathcal{L}[\phi_i, \partial_\mu\phi_i]$, $i = 1...N$, $\mu = 0, 1, 2, 3$, is a functional of N field variables $\phi_i(x)$ and their first (and possibily higher-order) derivatives. For the sake of exposition, I'll restrict attention to a single scalar field variable.

Typically, the functional integral over $\phi_H(x)$ in (1) is not exactly solvable, and even when it is, it may result in an effective Langrangian density that contains non-local terms (in the sense of depending on more than one spacetime point). These problems are jointly addressed by the second step in the construction of an EFT:

(II) The effective Lagrangian density is expanded in a local operator expansion

$$\mathcal{L}_{eff} = \mathcal{L}_0 + \sum_i c_i \mathcal{O}_i \tag{2}$$

where $\mathcal{L}_0$ can be taken to be the interaction-free Lagrangian density (for weak interactions), the c_i are coupling constants, and the sum runs over all local operators $\mathcal{O}_i$ allowed by the symmetries of $\mathcal{L}$.

Steps (I) and (II) can be characterized in the following ways:

(i) First, the effective Lagrangian density is formally distinct from the high-energy Lagrangian density. To the extent that this entails that the Euler-Lagrange equations of motion of the effective theory are distinct from those of the high-energy theory, the low-energy degrees of freedom $\phi_L(x)$ are dynamically distinct from the original degrees of freedom $\phi(x)$.[2]

(ii) Second, while the local operator expansion in Step II can be viewed formally as an approximate perturbative solution to the path integral (1), one can argue that an effective Lagrangian density is not simply an approximation of a high-energy Lagrangian density. In many cases, the exact form of the high-energy Lagrangian density is unknown, but an effective Langrangian density can still be constructed. Such a "bottom-up" EFT is obtained by including in the local operator expansion (2) all terms consistent with the symmetries and interactions assumed to be relevant at the energy scale of interest. A "folk theorem" identified by Weinberg (1979, pg. 329) then justifies viewing such bottom-up EFTs as not simply approximations to a high-energy theory.[3] This suggests that, even in the context of a "top-down" EFT for which a high-energy theory is known, the local operator expansion conceptually stands on its own.

(iii) Finally, the elimination of degrees of freedom in the construction of an EFT results from the imposition of a constraint (an energy, or minimum length, cut-off) directly on a Lagrangian density, as opposed to a set of equations of motion. Again, the result is a formally distinct effective Lagrangian density with a distinct set of equations of motion and a distinct set of dynamical variables.

---

[2] For a Lagrangian density $\mathcal{L}[\phi_i, \partial_\mu\phi_i]$, $i = 1...N$, the Euler-Lagrange equations of motion are defined by $\partial\mathcal{L}/\partial\phi_i - \partial_\mu(\partial\mathcal{L}/\partial(\partial_\mu\phi_i)) = 0$.

[3] The folk theorem states that "...if one writes down the most general possible Lagrangian, and then calculates matrix elements with this Lagrangian to any given order of perturbation theory, the result will simply be the most general possible *S*-matrix consistent with analyticity, perturbative unitarity, cluster decomposition, and the assumed symmetry principles" (Weinberg 1979, pg. 329).

3. An Interpretation of EFTs

The fact that EFTs come in two flavors, top-down and bottom-up, and that only the former is explicitly associated with a high-energy theory, might initially give one pause in attempting to formulate a notion of emergence appropriate for EFTs. In particular, the concern might be that such a notion assumes a distinction between a theory that describes emergent phenomena and a second theory that describes phenomena from which the former emerge; and such a distinction can only be made in the case of a top-down EFT. But this objection is easily blunted: Nothing in the construction of a bottom-up EFT precludes us from assuming that an associated high-energy theory exists; rather, the working assumption is simply that we do not know the form this high-energy theory takes. (A high-energy theory in this context need only be a theory that describes phenomena at an energy scale above that associated with an EFT; *i.e.*, it need not be a Grand Unified Theory applicable to all energy scales *in toto*.) Moreover, even in the top-down context, the EFT does not completely determine the form of the high-energy theory: for a given high-energy theory, more than one top-down EFT can be constructed.

These considerations suggest the following interpretation of EFTs, both top-down and bottom-up:

(a) *Failure of law-like deducibility*. If we understand the laws of a theory encoded in a Lagrangian density to be its Euler-Lagrange equations of motion, then the phenomena described by an EFT are not deducible consequences of the laws of a high-energy theory.

(b) *Ontological distinctness*. The degrees of freedom of an EFT characterize physical systems that are ontologically distinct from physical systems characterized by the degrees of freedom of a high-energy theory.

(c) *Ontological dependence*. Physical systems described by an EFT are ontologically dependent on physical systems described by a high-energy theory.

Claims (a) and (b) are suggested by the formal distinction between an effective Lagrangian density and a high-energy Lagrangian density, and their corresponding Euler-Lagrange equations of motion. In the case of (b), this suggests that the degrees of freedom of an EFT are dynamically distinct from those of a high-energy theory; moreover, the former are typically encoded in field variables that are formally distinct from those that encode the latter (*i.e.*, different field variables appear in the Lagrangian densities of an EFT and a high-energy theory). On the other hand, the fact that the degrees of freedom of the former can be identified, *via* Steps (I) and (II) outlined above, as the low-energy degrees of freedom of the latter suggests (c): the physical systems described by an EFT do not completely "float free" of the physical systems described by a high-energy theory.

I'd now like to flesh out the above interpretation with two examples, and then extract a notion of emergence from it. The following examples are of a top-down EFT for a 2-dimensional quantum Hall liquid, and a bottom-up EFT for general relativity.

Example 1. A Top-Down EFT for a 2-dim Quantum Hall Liquid.
The high-energy degrees of freedom of a quantum Hall liquid describe electrons moving in a 2-dimensional conductor and coupled to external magnetic and Chern-Simons fields. This is described by a non-relativistic Lagrangian density,

$$\mathcal{L} = i\psi^{\dagger}\{\partial_t - ie(A_0 - a_0)\}\psi - (1/2m)\psi^{\dagger}\{\partial_i + ie(A_i + a_i)\}^2\psi + \mu\psi^{\dagger}\psi + \vartheta\varepsilon^{\mu\nu\lambda}a_\mu\partial_\nu a_\lambda \qquad (3)$$

where the field variable ψ encodes the electron degrees of freedom, the pair (A_0, A_i), $i = 1, 2$, encodes the degrees of freedom of an external magnetic field, a_μ ($\mu = 0, 1, 2$) encodes the degrees of freedom of a Chern-Simons field, μ is the chemical potential, and the coefficient ϑ is chosen so that the electrons are coupled to an even number of "internal" magnetic fluxes, and hence refered to as "composite" electrons (Schakel 2008, pg. 349). Technically, this description entails that, in the presence of a strong external magnetic field, the electrons experience the quantum Hall effect. This occurs when the conductivity σ of the system becomes quantized in units of e^2/h; *i.e.*, $\sigma = \nu(e^2/h)$, where ν is called the "filling factor". The Integer Quantum Hall Effect (IQHE) occurs for integer values of ν and the Fractional Quantum Hall Effect (FQHE) occurs for values of ν given by simple fractions. Both the IQHE and the FQHE are characterized by incompressibility and disipationless transport, properties associated with superconductors. This suggests that these effects characterize a state of matter distinct from the conductor and refered to as a quantum Hall liquid.[4]

The properties of a quantum Hall liquid can be derived from the high-energy theory (3) by integrating out the electron degrees of freedom. The remaining degrees of freedom of the bulk liquid can then be identified with two Chern-Simons fields, a_μ, $(A_\mu + a_\mu)$, described by a "pure" Chern-Simons effective Lagrangian density,

$$\mathcal{L}_{eff} = \vartheta\varepsilon^{\mu\nu\lambda}a_\mu\partial_\nu a_\lambda + \vartheta'\varepsilon^{\mu\nu\lambda}(A_\mu + a_\mu)\partial_\nu(A_\lambda + a_\lambda) \qquad (4)$$

where the coefficient on the last Chern-Simons term is chosen to produce the integer QHE for the second CS field (Schakel 2008, pg. 349). This is an example of a topological quantum field theory (*i.e.*, a QFT encoded in a Lagrangian density in which a spacetime metric does not explicitly appear).

In this example, the high-energy Lagrangian density (3) is formally distinct from the effective Lagrangian density (4): (3) encodes a non-relativistic quantum field theory (QFT), whereas (4) encodes a topological QFT. This suggests that the laws of the EFT are not deducible consequences of the laws of the high-energy theory (*failure of law-like deducibility*); and that the EFT is dynamically distinct from the high-energy theory. Dynamical distinctness, coupled with the formal distinction between the field ψ that encodes the degrees of freedom of the high-energy

[4] The IQHE can be explained by reference to the discrete spacing between the energy levels of the system. The filling factor is given by ν = (#of electrons)/ (# of states per energy level). At integer values of ν, the first ν energy levels are full, and this entails incompressibility in the sense that no further electrons can be excited without a large cost in energy. The FQHE can be explained by noting that attaching an even number of fluxes to each electron cancels just enough of the external magnetic field to change the filling factor back to an integer value. Thus (in this description), the FQHE is the IQHE for composite electrons (Schakel 2008, pg. 343).

theory and the fields a_μ, $(A_\mu + a_\mu)$ that encode the degrees of freedom of the EFT, suggest that the later characterizes physical systems (*i.e.*, two topological Chern-Simons fields) that are *ontologically distinct* from those characterized by the former (*i.e.*, non-relativistic composite electrons). Finally, the fact that the degrees of freedom of (4) are exactly the low-energy degrees of freedom of (3) suggests that the physical systems described by (4) are *ontologically dependent* on those characerized by (3). In particular, the bulk quantum Hall liquid characterized by the topological fields a_μ, $(A_\mu + a_\mu)$ ultimately consists of non-relativistic composite electrons.

Example 2. A "Bottom-Up" EFT for General Relativity.
Recall that a bottom-up EFT is constructed in the absence of a high-energy theory by first identifying the relevant symmetries of the phenomenon in question and then constructing an effective Lagrangian density as a local operator expansion (2) that includes all possible interactions consistent with these symmetries. In the case of general relativity, these symmetries are general covariance and local Lorentz invariance. If one assumes that the metric $g_{\mu\nu}$ encodes low-energy degrees of freedom of an unkown high-energy theory, then an effective Lagrangian density corresponding to (2) can be given by,

$$\mathcal{L}_{eff} = \sqrt{g}\{\lambda + c_1 R + c_2 R^2 + c_3 R_{\mu\nu} R^{\mu\nu} + \ldots + \mathcal{L}_{matter}\} \tag{5}$$

where $g = \det(g_{\mu\nu})$, R, $R_{\mu\nu}$ are the Ricci scalar and Ricci tensor, the c_i are coupling constants, and the elipses refer to higher-order terms (Donoghue 1995, pg. 7). The Euler-Lagrange equations of motion generated by the first two terms are the Einstein equations with cosmological constant λ, and one can argue that the effect of higher-order terms is beyond current tests of general relativity.

In this example, since a high-energy theory is not known, the EFT is trivially characterized by the failure of law-like deducibility and ontological distinctness. Ontological dependence is secured by the assumption that the field variable $g_{\mu\nu}$ encodes the low-energy degrees of freedom of the unknown high-energy theory.

4. Emergence in EFTs

The philosophical literature typically distinguishes between two senses of emergence. The first views emergence as descriptive of the ontology (*i.e.*, entities or properties) associated with a physical system with respect to another. To say phenomena associated with an EFT are emergent in this ontological sense is to say the entities or properties described by the EFT emerge from those described by a high-energy theory. A second sense of emergence views it as a formal relation between theories. To say phenomena associated with an EFT are emergent in this sense is to say the EFT stands in a certain formal relation to a high-energy theory.

Note that an EFT does not stand in a precise mathematical relation to a high-energy theory. As outlined in Section 2, Step (I) in the construction of an EFT requires both a choice of cutoff and a choice of low-energy degrees of freedom with respect to the latter. These choices typically will be dictated by the specific context of the problem at hand, as opposed to being products of a

formal procedure. Similarly, the local operator expansion in Step (II) requires a context-specific identification of the symmetries of the high-energy theory (when it exists) or of the phenomena under investigation. This suggests that a purely formal concept of emergence for EFTs may not be appropriate. The approach adopted in this section will be to extract an ontological concept of emergence from the interpretation of EFTs suggested in Section 3. This interpretation motivates the following *desiderata*.

(i) First, the emergent system should ultimately be composed of microphysical systems that comprise the fundamental system and that obey the fundamental system's laws.

(ii) Second, the properties of the emergent system should not be deducible from the properties of the fundamental system.

I will follow Mainwood (2006, pg. 20) in refering to these *desiderata* as *microphysicalism* and *novelty*, respectively. They are underwritten in the EFT context by the elimination of degrees of freedom in the construction of an EFT. In particular, one might tell the following story about how the properties (and/or entities) of a system described by an EFT, encoded in an effective Lagrangian density $\mathcal{L}_{eff}$, emerge from a fundamental system described by a high-energy theory encoded in a Lagrangian density $\mathcal{L}$:

(i) First, the high-energy degrees of freedom are identified and integrated out of $\mathcal{L}$. This entails that the degrees of freedom of $\mathcal{L}_{eff}$ are exactly the low-energy degrees of freedom of $\mathcal{L}$. Thus is *microphysicalism* secured.

(ii) Second, the elimination of degrees of freedom also entails that the solution $\mathcal{L}_{eff}$ of the path integral (1) is dynamically distinct from $\mathcal{L}$, and is a functional of field variables that do not appear in $\mathcal{L}$. Dynamical distinctness suggests a failure of law-like deducibility from $\mathcal{L}$ of the properties described by $\mathcal{L}_{eff}$, and a difference in field variables suggests the properties and entities described by $\mathcal{L}_{eff}$ and $\mathcal{L}$ are ontologically distinct. Thus is *novelty* secured.

5. Other Notions of Emergence

To further flesh out the above notion of emergence for EFTs, it will be helpful to compare it with other accounts in the philosophical literature.

5.1. "New Emergentism", Spontaneous Symmetry Breaking, and Universality.

Mainwood (2006, pg. 20) characterizes the "new emergentism" of prominent condensed matter physicists (*e.g.*, Anderson 1972, Laughlin and Pines 2000) in terms of *microphysicalism* and *novelty*, as described above, underwritten by a physical mechanism. According to Mainwood, the specification of the latter is essential to avoid trivializing the concept of emergence: "...emergent properties are not a panacea, to be appealed to whenever we are puzzled by the properties of large systems. In each case, we must produce a detailed physical mechanism for emergence, which rigorously explains the qualitative difference that we see with the microphysical" (pg. 284). Such a mechanism plays both an explanatory and a formal role. First, it explains how novelty arises: New Emergentists "...follow a strategy of first exhibiting

evidence for emergence: the novel and unexpected character of certain systemic properties, and only then presenting a physical process - a 'mechanism' - that explains how such novelty can arise" (pg. 87). Second, formally, it underwrites the elimination of degrees of freedom from a constituative system, resulting in a system characterized by fewer degrees of freedom and exhibiting emergent phenomena. For Mainwood, the physical mechanism of most interest that accomplishes these tasks is spontaneous symmetry breaking (SSB): "The claim of the New Emergentists is that in the phenomenon of symmetry-breaking we have a mechanism by which the set of 'good coordinates' of the whole can be entirely different from the sets of good coordinates which apply to the constituent parts when in isolation or in other wholes" (pg. 107). However, Mainwood is careful to note that, in addition to SSB, the New Emergentists identify other mechanisms including renormalization, the integer and fractional quatum Hall effects, and localization (pg. 93), as well as universality (pg. 116).

SSB is the mechanism associated with the Landau-Ginzburg theory of phase changes in condensed matter systems, and its extension by renormalization group (RG) techniques. These theoretical frameworks associate phases with internal orders characterized by symmetries, and phase transitions with symmetry breaking. In the RG approach, phase transitions are analyzed by observing the behavior of a theory as its parameters are rescaled. Such rescaling generates a flow in the theory's abstract parameter space. A fixed point of such a flow is a point at which the values of the parameters remain unchanged under further rescaling; *i.e.*, they become scale invariant. This occurs at a critical point corresponding to a phase transition. Thus phase transitions are characterized by scale independence: the properties associated with a phase transition are independent of the micro-scale properties of the system. In general, there can be many distinct RG flows that terminate at a given fixed point. A fixed point x thus defines a *universality class* insofar as the theory defined by x is independent of the microphysical details of any theory on an RG flow that terminates at x.

Both SSB and universality play essential roles in two other recent discussions of emergence in physics. These accounts view universality as underwriting the ontological non-reductivism they deem necessary in descriptions of emergent phenomena, but differ on the significance of SSB. On the one hand, Batterman (2011, pg. 1034) has suggested that the notion of a protectorate (*i.e.*, a universality class) underwrites a concept of emergence "...that goes beyond mere claims to the effect that symmetry breaking occurs." According to Batterman (2011, pg. 1038), "It seems hardly satisfactory to appeal to symmetry breaking as an organizing principle independent of microdetails when we have such a profoundly successful story about why the microdetails in fact are largely independent or irrelevant." On the other hand, Morrison (2012, pg. 157) focuses explicitly on SSB as essential to the concept of emergence: "Although the RG provides an explanatory framework that shows why microphysical details can be ignored, it does not give us the kind of physical dynamics required for the production of emergent phenomena. For that we need symmetry breaking and the accompanying phase transitions". Morrison (2012, pg. 147) moreover suggests that "understanding emergent phenomena in terms of symmetry breaking -- a structural dynamical feature of physical systems... -- clarifies both how and why emergent phenoemena are independent of any specific configuration of their microphysical base." To support this claim, Morrison (2012, pp. 153-155) discusses an example due to Weinberg (1986) in which the essential properties of a superconductor are derived, not from a theory of its

microconstituents (*i.e.*, Cooper pairs), but by imposing symmetry constraints directly on a Lagrangian density.

Weinberg's example is instructive in the context of this essay insofar as it is an example of a bottom-up EFT. This raises two questions: First, how are SSB and universality related to EFTs, and second, if we agree with the above authors in their insistence on identifying a mechanism to underwrite a nontrivial concept of emergence, what is the nature of this mechanism in the EFT context?

The answer to the first question is explicit in the two examples discussed in Section 3: neither involves SSB or universality, at least as the latter is usually defined. Example 1 involves a phase transition from a less ordered conductor state to a more ordered quantum Hall liquid state; however, the orders cannot be distinguished by their symmetries. Wenn (1995, 2004) has developed a theory of "topological orders" that characterize the states associated with quantum Hall liquids, and argues that such liquids cannot be described by the standard Landau-Ginzburg theory of phase changes governed by SSB.[5] Moreover, while quantum Hall liquids may be described in terms of a concept of universality, assumedly it will not involve the same technical description as that provided by the RG analysis of fixed points.[6] In this broader sense, SSB is sufficient, but not necessary for universality. Example 2 also is not characterized by SSB or universality. In general, while the expansion point in the local operator expansion (2) of an effective Lagrangian density is defined by a fixed point (and hence a universality class)[7], an EFT itself need not be identified with a fixed point, nor, necessarily, with a point on an RG flow that terminates at a fixed point. Both of the latter correspond to renormalizable theories, whereas EFTs in general need not be renormalizable.[8] This suggests that a concept of emergence based on universality is too narrow for the EFT context. (It also suggests that a concept of emergence based on universality will have to include as emergent those phenomena associated with renormalizable theories; in particular, all the phenomena associated with the Standard Model would count as emergent.)

A concept of emergence appropriate for EFTs should thus be broader than a concept underwritten by SSB and/or universality. In Section 4 I suggested that emergence in EFTs be

[5] Wenn (1995, pg. 408) observes that the ground-state degeneracy that characterizes a quatum Hall liquid is not a consequence of the symmetry of the corresponding Hamiltonian, but rather depends on spatial topology. This fact, together with the fact that the ground-state degeneracy is robust under perturbations, suggests to Wenn that it be associated with a notion of universality characterized, not by symmetries and RG fixed points, but by topological order.

[6] Mainwood (2006, pg. 264, f.n. 3) acknowledges that the general concept of a universality class as used by New Emergentists "...is clearly meant to also extend beyond areas in which the RG techniques are usually applied".

[7] For weak interactions, the point of expansion $\mathcal{L}_0$ is taken to be a Gaussian fixed point in the parameter space of the high-energy theory, but any fixed point will serve this purpose. In general, the terms in the local operator expansion (2) are characterized by their behavior with respect to a fixed point (they can either increase, decrease, or remain the same as the RG flow approaches the fixed point), and this characterization is essential to the behavior of the EFT.

[8] A fixed point corresponds to a renormalized theory; *i.e.*, a theory that is energy scale-independent. A point on an RG flow that terminates at a fixed point corresponds to a non-renormalized renormalizable theory; *i.e.*, a "bare" theory that is capable of being made energy scale-independent, but whose parameters have not yet been rescaled to make this so. The most general form (2) of an EFT encompasses both of these theory types, but also a theory represented by a point on an RG flow that passes through a neighborhood of a fixed point, but does not intersect it. Such a theory is non-renormalizable.

characterized in terms of microphysicalism and novelty, and that these characteristics are underwritten simply by the elimination of degrees of freedom in the construction of an EFT. Both Mainwood and Morrison require a causal/mechanical explanation of emergent phenomena in terms of a physical dynamical process like SSB (Batterman, on the other hand, is content with a unifying explanation based on the renormalization group). Morrison (2012, pg. 160), in particular, views an appeal to the elimination of degrees of freedom as not enough: "[t]he important issue...is not just the elimination of irrelevant degrees of freedom; rather it is the existence or emergence of cooperative behavior and the nature of the order parameter (associated with symmetry breaking) that characterizes the different kinds of systems." In response, I would agree that, by itself, an appeal to the elimination of degrees of freedom does not explain the existence of cooperative behavior, nor does it explain the existence and novel nature of emergent phenomena. On the other hand, the particular emergent phenomena associated with EFTs are not essentially characterized by cooperative behavior: while some examples are (quantum Hall liquids as described by EFTs), others are not (general relativity as described by an EFT). Moreover, within the interpretive framework suggested in Section 3, the elimination of degrees of freedom in an EFT does fulfill a causal/mechanistic explanatory role. In particular, the elimination of degrees of freedom in an EFT explains the existence and novel nature of low-energy emergent phenomena by explaining how they are related to high-energy phenomena by a failure of law-like deducibility, and by ontological distinctness tempered by ontological dependence. Thus I would argue that the particular type of elimination of degrees of freedom in an EFT, coupled with an appropriate intepretation of EFTs, succeeds in doing the explanatory work deemed necessary by Mainwood and Morrison for a nontrivial concept of emergence.

5.2. "Weak Ontological Emergence".

An approach to a concept of emergence that stresses the importance of the elimination of degrees of freedom is given by Wilson (2010), who refers to this concept as "weak ontological emergence".[9] The elimination of degrees of freedom in a theory in physics, according to Wilson, involves the imposition of constraints that eliminate functional dependences between system properties and some subset of degrees of freedom (pg. 284).[10] Wilson takes the following to be examples of this:

1. The electric field of a spherical conductor, which depends only on the degrees of freedom of the charges on its boundary (pp. 285-286).

2. Statistical mechanical aggregates: "[S]uccessful applications of the RG method to certain composed entities indicate that such entities have DOF [degrees of freedom] that are eliminated relative to systems consisting of their composing [parts]" (pg. 288).

3. Quantum degrees of freedom in the classical limit (pp. 288-290).

[9] Wilson (2010, pg. 280) takes "weak ontological emergence" to be compatible with physicalism, as opposed to "strong ontological emergence", which is not.

[10] Wilson (2010, pg. 282) considers a more general notion of a degree of freedom that the one adopted in Section 2, allowing that it need not necessarily figure into a state description that underwrites a dynamics.

These examples arise in different contexts, none of which is appropriate for EFTs. Example 1 arises in the context of a single theory by the imposition of boundary conditions on the theory's equations of motion; thus it does not apply to the EFT context which involves two formally and dynamically distinct theories. Example 2 is drawn from discussions in Batterman (2002) and elsewhere and arises in the context of two theories (statistical mechanics and thermodynamics) related by a limiting relation. Arguably, this example also does not apply in general to the EFT context: Briefly, the procedure involved in constructing an EFT, as outlined in Section 2 above, does not produce a limiting relation between the EFT and its high-energy theory (see Bain 2012, pp. 28-32, for further discussion of Batterman's (2002) notion of emergence in the context of EFTs). Finally, Example 3 also seems to arise from an assumed limiting relation between two theories (classical and quantum mechanics), and thus is not applicable to EFTs. (The nature of the limiting relation in Example 3 is a bit more controversial than in Example 2, insofar as more than one dynamically distinct quantization of a given classical system can be constructed).

In the construction of an EFT, the elimination of degrees of freedom is not characterized by a limiting relation between theories, nor by the imposition of constraints on a set of equations of motion. Rather, it is characterized by the imposition of a constraint (in the form of a boundary condition that imposes an energy, or minimum length, cutoff) directly on the degrees of freedom of a Lagrangian density, as opposed to its equations of motion. This yields a formally distinct effective Lagrangian density with a distinct set of equations of motion. This formal distinctness severs functional dependences between the remaining low-energy degrees of freedom and the dynamics of the high-energy theory.

This type of elimination of degrees of freedom in an EFT does not appear to be what Wilson has in mind. Wilson takes the sort of elimination of degrees of freedom that underwrites ("weak ontological") emergence to play two roles. First, it establishes the physical acceptability of an emergent entity by securing the law-like deducibility of its behavior from its composing parts. This is taken to partially underwrite a concept of physicalism:[11]

> ...so long as a given special science treats only of entities E whose characterization requires the same or fewer DOF [degrees of freedom] as their composing e_i, the special science is appropriately seen as extracted from the more fundamental science treating the e_i, such that the laws of the special science (expressing, in particular, the properties and behavior of E) are deducible consequences of the laws of the more fundamental science (expressing, in particular, the properties and behavior of the e_i). This is the case, in particular, with the special sciences (statistical and classical mechanics) treating entities satisfying *Weak ontological emergence* (Wilson 2010, pg. 295).

[11] For Wilson, physicalism in the context of weak ontological emergence is also underwritten by the claim that "...the law-governed properties and behavior of [an emergent entity] are completely determined by the law-governed properties and behavior of the [composing entities]..." (2010, pg. 280). If "completely determined" refers to an ontological notion of dependence between the emergent and fundamental entities, then this amounts to the notion of microphysicalism in Section 4. But if "completely determined" refers to a formal characteristic of a set of equations of motion, then I would argue that it is too strong a criterion on which to base a notion of physicalism. In particular, it fails in the context of typical EFTs.

Second, according to Wilson, the elimination of degrees of freedom entails that an emergent entity is characterized by different law-governed properties and behavior than those of its composing parts. This is taken to underwrite a failure of ontological reductionism:

> The line of thought appeals to the laws that scientists take to govern an entity of a given type, as providing an appropriate basis for identifying the DOF associated with that entity... [The argument] concludes that [the emergent entity] E is not identical to [its composing parts] e_r, on grounds that there are scientific reasons for associating E with certain laws, such that specifying E's law-governed properties and behavior requires certain DOF; and for associating e_r with certain laws, such that specifying e_r's law-governed properties and behavior requires certain DOF *different* from those required to characterize E (Wilson 2010, pg. 301).

This failure of ontological reductionism might charitably be associated with a notion of novelty, and this, coupled with physicalism might suggest a similarity between Wilson's weak onotological emergence and the sense of emergence in EFTs expounded in Section 4 above. However, again, the elimination of degrees of freedom that underwrites Wilson's physicalism and the failure of ontological reductionism is decidedly different from that which underwrites microphysicalism and novelty in EFTs: Where Wilson suggests elimination of degrees of freedom *secures* the law-like deducibility of an emergent entity from its composing parts, I've suggested that elimination of degrees of freedom in an EFT is characterized, in part, by a *failure* of law-like deducibility, and take this to underwrite novelty (in the sense of dynamical and ontological distinctness). I've also suggested that elimination of degrees of freedom in an EFT is also characterized by the retention, in the EFT, of the low-energy degrees of freedom of the high-energy theory, and it is *this* fact that underwrites a concept of (micro)physicalism (as opposed to a relation of law-like deducibility). Thus, while Wilson's concept of emergence may be applicable to some subset of physical systems described by theories in physics, it is not applicable to EFTs, under the interpretation suggested in Section 3.

6. Conclusion

This essay suggests that emergence in an EFT can be characterized by novelty and microphysicalism underwritten by the elimination of degrees of freedom from a high-energy theory. This is an elimination of degrees of freedom imposed directly on a high-energy Lagrangian density, as opposed to a set of equations of motion. It results in an effective Lagrangian density that can be interpreted as describing novel phenomena in the sense of being dynamically independent of, and thus not deducible from, the phenomena associated with a high-energy theory. These novel phenomena can be said to ultimately be composed of the phenomena that are constitutive of a high-energy theory, insofar as the degrees of freedom exhibited by the former are exactly the low-energy degrees of freedom exhibited by the latter. Finally it was argued in Section 5 that this concept of emergence in EFTs is more general than concepts of emergence based on spontaneous symmetry breaking and/or universality, but more narrow that a concept based simply on the elimination of degrees of freedom.

References

Anderson, Philip. 1972. "More is Different." *Science* 177: 393-396.

Bain, Jonathan. 2012. "Effective Field Theories." Forthcoming in *The Oxford Handbook of Philosophy of Physics*, ed. Robert Batterman. Oxford: Oxford University Press.

Batterman, Robert. 2002. *The Devil in the Details*. Oxford: Oxford University Press.

_____. 2011. "Emergence, Singularities, and Symmetry Breaking." *Foundations of Physics* 41: 1031-1050.

Donoghue, J. F. (1995). "Introduction to the Effective Field Theory Description of Gravity." arXiv:gr-qc/9512024v1.

Laughlin, Robert, and David Pines. 2000. "The Theory of Everything." *Proceedings of the National Academy of Sciences* 97: 28-31.

Mainwood, Paul. 2006. "Is More Different? Emergent Properties in Physics." PhD diss., University of Oxford. Online at PhilSci Archive: <http://philsci-archive.pitt.edu/8339>.

Morrison, Margaret. 2012. "Emergent Physics and Micro-Ontology". *Philosophy of Science* 79: 141-166.

Schakel, Adriaan. 2008. *Boulevard of Broken Symmetries: Effective Field Theories of Condensed Matter*. Singapore: World Scientific.

Wen, Xiao-Gang. 1995. "Topological Orders and Edge Excitations in Fractional Quantum Hall States". *Advances in Physics* 44: 405-473.

_____. 2004. *Quantum Field Theory of Many Body Systems*. Oxford: Oxford University Press.

Weinberg, Steven. 1979. "Phenomenological Lagrangians." *Physica* 96A: 327-340.

_____. 1986. "Superconductivity for Particular Theorists." *Progress of Theoretical Physics Suppplement* 86: 43-53.

Wilson, Jessica. 2010. "Non-reductive Physicalism and Degrees of Freedom." *British Journal for Philosophy of Science* 61: 279-311.

Deep Conventionalism about Evolutionary Groups

Matthew J. Barker, Concordia University

Joel D. Velasco, California Institute of Technology

Concepts of evolutionary groups are some of the most important concepts in biology and its philosophy. These groups include often-cited players in evolutionary processes, such as populations, species, biological races, and lineages of various sorts. In a broad sense, certain products of evolution are also considered evolutionary groups, including clades of species, of populations, of organisms, and of gene families. Assumptions about evolutionary groups feature in nearly every biological study, whether explicitly evolutionary, molecular, or otherwise. And philosophers have exported views about evolutionary groups as far afield as debates about how we should organize and fund science in democratic societies.[1]

The widespread importance of concepts of evolutionary groups helps make disputes about them important. But it makes perhaps even more important a rare consensus. The consensus is a form of objectivism about what determines which collections are evolutionary groups. It allows that our research interests may help determine which group concept is best in a given case. But it says that on any single prevailing group concept, we as minded agents do not help fix or determine which candidate groups are indeed evolutionary groups under that concept; instead, objective facts alone suffice for that. Although a mix of biological, chemical, psychological, physical facts and so on may be recognized among these objective facts, it is harmless in this

context to use "biological facts" to refer them. What is important is that this set includes only mind independent facts. It excludes by definition those facts that are instead fully or partially mind dependent, e.g., facts about our research interests, perceptual abilities, general values, and so on. Given these terms, the consensus view says objective biological facts alone suffice—we as minded agents are not needed—to determine whether a candidate group is in fact a kind of evolutionary group.

Explicit statements of objectivism about evolutionary groups in biology literatures are typically each about one or another specific kind of evolutionary group. And fellow biologists seldom challenge these. When molecular phylogeneticists and developmental botanists argue that the *AGL6-like* family of genes is a clade that has existed for at least 300 million years, colleagues may dispute whether the *AGL6-like* group really is a clade.[2] But the vast majority on either side of any such dispute will agree that it is the biological facts alone that determine whether the *AGL6-like* group satisfies the notion of clade that they all (let us suppose) are using. In another chapter of the objectivist consensus, evolutionary ecologists argue that many a biological taxon has objective cohesion owing to gene flow between but not beyond the populations constituting it.[3] Again, any disputes about this will very probably not indict objectivism. Indeed, objectivism about evolutionary groups is typically taken for granted *without* explicit statement. And when stated, authors happily leave it as an assumption.[4] What could be more obvious than, say, that a clade of plants would be a clade even were we never here to discover that?

Philosophers have more explicitly treated or adopted objectivism about evolutionary groups as a *general* consensus, rather than dwelling only on more specific objectivisms about this or that kind of group. For example, Dupré, Ereshefsky, and Kitcher clarify that their respective pluralisms about biological classification are consistent with objectivism about many kinds of evolutionary groups (though they may disagree on some kinds of groups).[5] But their discussions do not aim for, and so understandably do not provide, close scrutiny or detailed defense of objectivism about evolutionary groups. The basic and assumed idea is that many different evolutionary groups are, despite their differences, similarly objective because the evolutionary processes that involve and produce such groups operate objectively.

The sway the consensus holds in both local chapters and as a whole is remarkable. Objectivism about clades lies behind the common view that there is a single universal tree of life. Objectivism about taxonomic groups prevails among even non-objectivists about taxonomic ranks, and is part of the idea that any one species concept univocally classifies organisms (barring vagueness) despite competing species concepts ambiguously cross-classifying them.[6] Authors working on the Human Genome Diversity Project have used population objectivism to justify decisions about what kind of informed consent to acquire and when, and about which research methods and data to use.[7] And the objectivist consensus has motivated attempts in more general philosophy of science to retain a form of scientific realism despite recognizing an increasing number of ways in which values (in a general sense) must shape scientific inquiry.[8]

Despite its dominance, we will argue that this consensus is mistaken because objectivism about many and perhaps all commonly recognized kinds of evolutionary groups is mistaken. This paper aims to displace the consensus with a new view, Deep Conventionalism.

This new view consists of two parts. The first is a pluralism that is deeper than familiar pluralist views attributed to Dupré, Kitcher, and Ereshefsky. Unlike their pluralisms, ours undermines the objectivism of the consensus. The second part of our view fills this void with a conventionalism that applies to a wide variety of evolutionary groups. This conventionalism says that even given any single, specific evolutionary grouping concept, typically something more than the objective biological facts must determine or fix which things are such groups. The "something more" is a mix of facts about us. The mix includes various conventions of ours, but also our research interests, values, abilities, and so on. We use "conventionalism" for short.

To proceed, we first situate Deep Conventionalism among related views. This positions us to clarify a key notion of *suppressed variables* and the deep pluralism associated with these. We then undertake the central task of showing how such variables ensure that our view holds for a variety of evolutionary grouping concepts, using cohesive functional units, populations, and clades as exemplars. Finally, we

discuss potential objections and highlight implications for a range of important positions.

1. Situating Deep Conventionalism

What is an evolutionary group? An innocuous answer is that an evolutionary group is any group of things that have certain evolutionarily salient relations that set them apart from other things. Exactly when things enjoy such relations, they make an evolutionary group out of what would otherwise have been, from an evolutionary perspective, a mere group or collection.

The category of "evolutionary group" divides into distinct kinds of evolutionary groups. Authors recognize these with definitions and elaborations of different evolutionary grouping concepts, distinguished by appeal to different evolutionarily salient relations. For instance, "species" is said to name one kind of evolutionary group, and "population" another. Definitions of the species concept typically attempt to identify the evolutionarily salient relations between organisms and/or groups of them in virtue of which those things together form a species. Definitions of "population" typically attempt to identify the relations in virtue of which organisms form a population.

Sometimes a dispute about a group concept is about which definition of it is "correct", "best" or "legitimate"—about which identified relations are the ones that make a group a species, or a population, etc. This is sometimes further understood as

competition between more specific concepts, each vying to be the specification of the broader concept under dispute, the one that objectively does or does not apply to each candidate group. Someone with this understanding sees the biological species concept, phylogenetic species concepts, and ecological species concepts battling to be *the* objective species concept.

What we will call *pluralistic objectivism* is an increasingly popular way of interpreting these sorts of disputes differently, a way of qualifying or even eliminating some of the dispute in each case while still conforming to the objectivist consensus.[9] This pluralism and how it differs from our own is most easily seen by drawing from its application in the species concepts literature, though when later arguing for our own view we will only treat species concepts implicitly, through discussions of other related important kinds of evolutionary groups.

Pluralistic objectivism has two noteworthy features. First, it claims that the concept in question either subsumes or should be eliminated in favor of two or more—a plurality—of finer-grained concepts, each of which is legitimate. This pluralism about legitimacy allows that two concepts thought to be competing for legitimacy are instead each legitimate for distinct purposes. For example, the biological species concept is said to be legitimate (or the best, or the correct concept to use) for some purposes, phylogenetic species concepts for others, and ecological ones for still others. Our interests help determine which finer-grained concept of a given broader type is best in which case.

Second, pluralistic objectivism says that for each finer-grained concept it is objective facts alone that suffice to determine how a set of things divides into groups under that concept. So our interests help determine which of these concepts to use in a given case, but the objective facts have already determined what the groups are under each of the concepts our interests are choosing between.

More precisely, take the set of organisms, *S*, consisting of two populations of organism on opposites sides of a mountain, population North and population South. According to pluralistic objectivism, objective facts suffice to determine whether and how *S* divides into biological species groups, or just biological *groups* when these are deemed objective but their assignment to the species rank is not, and whether and how it divides into ecological species groups, or just ecological *groups*. Suppose the pluralistic objectivist believes that objective facts determine that North and South together form one biological species group, while objective facts also determine that North is one ecological species group, and South another. That is, one finer-grained concept lumps North and South, the other splits them. Then, in (say) a research study or classification project involving *S*, our interests enter the picture, helping determine whether it is best or legitimate for us to recognize the lumping divisions of the biological species concept, or the splitting divisions of ecological ones, or both, or neither. Independently of us, the divisions are there.

In typical empirical conditions our view disagrees with the objectivism in pluralistic objectivism. There are many ways to convey this, with some ways useful to some people and others to others. Our view says that conventions—facts about our interests, values, abilities, and so on—help determine not only which concept is relevant or legitimate in a given case, but also to which candidate groups it applies in that case. Our conventions are needed along with objective facts to fix the extensions of evolutionary group concepts. Conventions are needed to determine whether a collection of organisms is a population, clade, give type of species, etc. How many biological species are in a given set of organisms is not fully fixed by objective facts across all research contexts; in some research contexts facts about us pair with objective facts to give one count, in other research contexts they pair to give others. Conventions help determine not only significance but also accuracy of group identify (and associated taxonomic) claims. In typical empirical conditions, the biological facts cannot determine whether North and South form distinct ecological species, let alone species simpliciter.

As we will clarify near the end of the paper when discussing concept splitters, our conventionalism about grouping concepts, including finer-grained ones, should not motivate pluralistic objectivists to simply split their finer-grained concepts into even finer-grained ones in the hope of reaffirming objectivity at still finer conceptual levels. We are stuck with our conventionalism and should abandon pluralistic objectivism. Our reasoning begins in the next section by clarifying what we call Indeterminacy Pluralism. This is pluralism with respect to the values that can be

taken by the *suppressed variables* associated with any single prevailing evolutionary grouping concept, not pluralism about multiple concepts being legitimate. To understand suppressed variables, we start with a non-biological, linguistic example. But we stress that this is only to intuitively convey the *form* that Indeterminacy Pluralism takes, and how it can mandate conventionalism. We will then have to show that the biological cases take this form. Distant views in philosophy of language do no work in any of this.

2. Suppressed Variables

Suppose Charles is at a large picnic with much of Alfred's extended family. Alfred is in a small group of people around a punch bowl, and Charles, walking towards them, senses that the small group is not enjoying the live country music. But the rest of the people at the picnic love the music. Charles asks, "So is this small group of you unified in your response to country music?"

Alfred answers "yes." But this is correct only by drawing on context to further specify the question. Alfred gathers that Charles asked his question with certain *kinds of responses* in mind, and certain *kinds of country music*. Without explicitly or implicitly choosing particular values for these variables, there is no correct answer to the question. And on other values of the variables, we can imagine that the relevant facts ensure that Alfred's answer is instead *not* correct.

Take the case in which the small punch bowl group includes just Alfred and his brother and sister. For the *kinds of response* variable, choose the "emotional response" value. For the *kinds of country music* variable, choose the "pop-country music" value. Then, given facts about his family, Alfred can assure you that he was correct to affirm that the small punch bowl group is unified in its response to country music. He and his siblings each react with disgust to pop-country music, and more so than any of the attending extended family does. However, now change the value of the *kinds of country music* variable to "alt-country music." Then Alfred's affirmative answer to Charles's question switches to *not* correct. Alfred likes alt-country music and his brother loves it. But his sister detests it, more than any people in the extended family. Changing the other variable, from "emotional response" to "sensory-motor response," may also make Alfred's affirmative answer incorrect.

In cases like the picnic scenario, semantic facts about the meaning of "response to country music" leave many variables open. Short of further inputs, there is no semantic fact of the matter about whether the *kinds of response* variable takes the "emotional response" value or "sensory-motor response" value. Given that such variables *do* often get fixed in the face of these factual shortfalls, something else must add to the semantic facts to fix the variable values.

In the picnic case, that "something else" is pretty clearly our conventions about contextual information. Suppose that at the picnic it is pop-country music, in particular, that is playing when Charles asks his question. Then very probably, both

he and Alfred have in mind the "pop-country music" value of the *kinds of country music* variable. And this is most likely because both of them are following a reasonable convention, which here implies: if it is pop-country music that is playing at the picnic, then presume that the kind of country music that the question is about is pop-country music. Indeed, it seems that in cases with conditions like this case, conventions must help with any fixing of variable values.

The relevant biological variables, not just linguistic ones, are also of this kind and lead to similar results. To see this, first consider that in the picnic case we have Indeterminacy Pluralism consisting in two conditions. One: whether a group of people is unified in its response to country music depends on variables that can each take one of a plurality of values that are all included among the facts. In fact, Alfred emotionally responds to alt-country music in one way, and to pop-country in another. Two: for some or all of these variables some different available values would on their own lead to incompatible results, e.g., to the punch bowl group having a unified response on some variable values but not on others. So the facts independent of our contributions leave it indeterminate whether the punch bowl group is unified in its response to country music. Given that indeterminacy in some cases like this *is* overcome, our contributions are needed to make up those indeterminacy shortfalls.

Analogously for prevailing kinds of evolutionary groups, Indeterminacy Pluralism is true and concerns the plurality of values that are available for variables of being an evolutionary group of the given kind. Regardless of whether there is a plurality of

legitimate species concepts as familiar pluralisms claim, the above two conditions are typically met when using any one of these or any other prevailing evolutionary group concepts. And again we must make up this shortfall conventionally. To make good on these claims, the next three sections discuss prominent examples of forward looking evolutionary groups and then backward looking evolutionary groups; for brevity, general objections are discussed after these sections rather than repeated in each.

3. Functional Units and Cohesion

Many evolutionary groups are what Baum calls "functional units", characterized by "cohesion or causal efficacy" that allows them to be "players" or forward looking groups in ongoing evolutionary processes (*op. cit.*, p. 74). Although authors, including Baum, typically have species in mind when discussing these units, some note that the cohesion that is supposed to make species functional units is also had to greater degrees by some non-species groups, such as populations, and to lesser degrees by other non-species groups, such as multi-species syngameons and perhaps some higher taxa.[10] We dwell first on the species grade of this cohesion: species cohesion.

Species cohesion has been important in many articulations of the nature of species since the Modern Synthesis.[11] This is explicit in some species concepts, such as the evolutionary species concepts of Wiley and Simpson, and implicit in others, such as Mayr's biological species concept.[12] Species cohesion is also important to various interventional and field studies, e.g., attempts to explain why conspecific populations together trace a distinct trajectory through the space of evolutionary pressures,

including various forms of natural selection. Some such projects attempt to discover and mathematically represent relationships between effective population sizes, population subdivision, migration, and species cohesion. For instance, a traditionally recognized relationship is that the effective number of migrants, N_em, from one population to another must be ≥ 1 for "maintaining species cohesion" across those populations.[13] Studies of evolutionary forces attempt to refine this view.[14] Although the importance of species cohesion and similar sorts of functional cohesion differ from the importance of the clades in phylogenetics, many phylogeneticists insist that species are special precisely because of their functional cohesion.[15]

The question for us is whether species cohesion is a conventional sort of unity due to featuring suppressed variables. Only recently have authors provided the clarification of "species cohesion" required to answer this.[16] Species cohesion is a grade of evolutionary response cohesion that involves organisms or populations responding similarly to evolutionary pressures. Importantly, whether a group responds in such a way depends partially on the contrast class. Take a collection of populations. It manifests evolutionary response cohesion exactly when the responses of its populations to evolutionary pressures are more similar to each other than to any outside the collection. This is for a collection to be exclusive, in at least one way, among others. Without this particular relativization to things outside the collection, it is hard to see how the collection could have the cohesion that is supposed to set it apart from other things – give it functional *unity*.

Once it is clear that evolutionary response cohesion distinguishes evolutionary groups that we call functional units, it is easy to see that being such a unit depends on the values that suppressed variables take. These are variables of evolutionary response cohesion. Recall populations North and South, flanking the mountain. They will face many evolutionary pressures, often concurrently: a drought, a nutrient deficiency, emergence of an advantageous mutation. And there are different responses they can have to any one pressure: *this* trait declines in frequency in one population and increases in the other; *that* trait increases in both populations. Minimally then, two suppressed variables of evolutionary response cohesion (of any grade) that can take many values are *which evolutionary pressures* and *which aspects of response*.

In typical cases, there will be an enormous number of values these variables can take because organisms and populations have many traits and face many evolutionary pressures. On many combinations of these values the two mountain populations would count as having evolutionary response cohesion while on many others, they would not. Suppose that in each population, just 1% of organisms have a suite of genes that, during depressed humidity, contribute to their retaining moisture far better than the other 99% of organisms. Then there is a series of devastating droughts. The suite of genes increases to 35% representation in both populations. In organisms of other nearby populations, genes involved in moisture retention are quite variable, resulting in no pattern of frequency response during the droughts. Choosing “moisture retention genes” for the *which aspects of response* variable, and

"series of droughts" for the *which evolutionary pressures* variable, along with many other values of these variables that similarly relate the populations, would count the two mountain populations as having associated evolutionary response cohesion. The responses of moisture retention genes in those two populations are more similar to each other than to any responses in other populations.

At the same time, in North, and in all populations nearby *except* South, a new sequence at a genetic locus has emerged that dramatically helps utilize increased sunlight hours for energy production. Spikes in sunlight hours accompany the droughts. Selection then facilitates a spike in population frequencies of the new sunlight utilization sequence—except in South, which does not yet enjoy that sequence. If we change the value of the *which aspects of response* variable, from "water retention genes," to "sunlight utilization locus" plus other aspects of response that similarly relate all the populations, then the two mountain populations would not count as having evolutionary response cohesion.

This clarifies how functional units distinguished by evolutionary response cohesion will typically satisfy the two conditions of Indeterminacy Pluralism. To help verify that this is typically so, most any study of population differentiation will do. Barbará et al. (*op. cit.*) recently described a nice model for studying population differentiation across continental radiations. The model involves populations of *Alcantarea* species, perennial plants in Brazil that grow on large granite outcrops (similar to Ayers Rock, aka Uluru). Populations in these species made a useful model partly because

measurements suggested that factors known to complicate some population differentiation studies (e.g., populations diverging markedly from Hardy-Weinberg and selection/drift equilibriums) were absent, or otherwise would not significantly distort assessments of these populations.

Highly varied traits characterized organisms in these populations. For example, all eight microsatellite loci investigated in populations of one species, *Alcantarea imperialis*, "were polymorphic, with up to 14 alleles per locus" (*ibid.*, p. 1985). And the scattering of populations across granite outcrops suggests varied evolutionary pressures across those populations. Together these points indicate there are many values that the variables *responses to evolutionary pressures* and *which aspects of response* will take across the studied populations of *Alcantarea imperialis* (first condition of Indeterminacy Pluralism). Also, evidence suggested that for at least some of these variables some different available values would on their own lead to incompatible verdicts on whether the populations of the *Alcantarea imperialis* jointly manifest the species grade of evolutionary response cohesion (second condition of Indeterminacy Pluralism). Genetic distances between populations of *Alcantarea imperialis*, for example, were sometimes nearly as large as between that species and another *Alcantarea* species (*ibid.*, p. 1986). Genetic variance, too, between conspecific populations was near what it was between the species (*ibid.*, p. 1988), and many researchers believe that in many cases variance between conspecific populations is even greater than that between species. These *statistical* measures of distance and variance strongly suggest that many *particular* genetic responses to evolutionary

pressures are more similar between populations of distinct species than between conspecific populations.

Generally across functional unit candidates, many of the biological values available for suppressed variables of evolutionary response cohesion would count the candidate as being a functional unit. Many other available biological values would have the opposite result. Both results cannot obtain. And the biological facts do not choose which of all the biological values are taken by the variables. We must do that. Species cohesion and other grades of evolutionary response cohesion are therefore conventional sorts of unity in light of the Indeterminacy Pluralism that is true of them. This entails conventionalism about functional units distinguished by such cohesion.

4. Populations and Interaction Rate Exclusivity

Not all forward looking functional units are distinguished by some grade of evolutionary response cohesion. For others, it is how they causally interact with each other, rather than how they causally respond to shared evolutionary pressures, that makes them functional units of an evolutionary kind.[17] Populations are the prime example.

Millstein usefully compares prevailing distinct population concepts in terms of permissiveness.[18] Some are astonishingly permissive, recognizing any collection of organisms within a species as a population (*ibid.*, p. 61). For our purposes it would be

most convincing to show that the least permissive, or most specific, population concept that is common in evolutionary studies features Indeterminacy Pluralism. Millstein, following in the wake of others, refines the definition of such a concept. Roughly, "the causal interactionist population concept" says that a population is any group of multiple conspecific organisms that is the largest group for which the internal rates of survival and reproduction interactions are much higher within the group than outside it (*ibid.*, p. 67).

As with evolutionary response cohesion, the evolutionary group-making property that this definition picks out is a kind of unity or exclusivity property. It is relativized to things outside candidate populations, as you would expect of a property that is supposed to unify and set apart a group from other things. In this case, it is survival and reproduction interaction rates that are supposed to be distinctive between group members, relative to outsiders. Effectively these interaction rates are to be greater between group members than between them and outsiders.

This property also features Indeterminacy Pluralism due to variables that can take many values, some large sets of which would suggest a group has the property and other large sets of which would imply otherwise. We find these variables at more than one level. At a first level, there is a variable that is not suppressed at all, the *kind of interaction variable*. It is not suppressed because two values of this variable – "survival interaction" and "reproduction interaction" – are explicitly referenced in the description of the definitive property. These two values can pull in opposite

directions. Many organisms frequently interact with others in a way that changes their life expectancy (e.g., negatively in the case of direct or indirect competition, and positively in the case of cooperation), without changing their expected reproductive output (*ibid.*). The situation escalates if we omit the stipulated restriction of a population to members of the same species, as Godfrey-Smith suggests we do to properly understand natural selection, and as one must (on pain of circularity) if one defines "species" in terms of populations.[19] Highest rates of reproductive interactions for some plant in my garden might connect it with pollinators and seed dispersers, while highest rates of survival interactions might connect it with other plants crowding it for soil and sun.

One level down we find two suppressed variables: *kinds of survival interaction* and *kinds of reproductive interaction*. These can take several values, indicated when Millstein notes there are several different kinds of survival and reproductive interactions, respectively (*op. cit.*, pp. 67-68). Among the reproductive kind, she cites successful matings, unsuccessful matings, and different offspring rearing activities. Survival interactions include direct competition, indirect competition, and cooperation. Values for each of these will often simultaneously pull in opposite directions with respect to a candidate group's being "interaction rate exclusive." A tree in Mauro's backyard has perennially poor fruit. The local deer nearly always choose the neighbor's tree fruit instead. Furthermore, most of the fruit from Mauro's tree rots below it, leaving seeds to struggle for the little light penetrating through other crowding trees of Mauro's. The struggling seeds of Mauro's tree involve that

19

tree in frequent (unfavorable) reproductive interactions with Mauro's other trees; the fruit of that tree involve it in frequent (unfavorable) reproductive interactions with the neighbor's tree. Many organisms frequently each have many reproductive interactions, some of which suggest connections to one group, some to another, others to another still, and so on. Likewise for their survival interactions.

Suppose we accept that for many a candidate population in the popular sense that Millstein refines, many values for the variables we have discussed would imply that the candidate has the exclusivity or unity that marks such populations. And many other values would imply the candidate does not have this property. Then we again have Indeterminacy Pluralism, and many population boundaries must be ones we help fix. Populations popularly conceived are then conventional in our sense.

5. Clades, Splitting and Genealogical Exclusivity

In many areas of biology the central evolutionary grouping concept is that of a clade or a monophyletic group. Clades are evolutionary groups because they feature a kind of evolutionary unity - they are united by a shared common ancestry, which makes them backward looking groups. Relative recency of common ancestry often explains why members of a clade share the traits that they do, grounds a variety of inferences about the past, and provides evidence about what unseen traits in members of the group will be like. Such features make clades so important in taxonomy that a common view of biological taxa is that they *must* be clades. The importance extends far beyond taxonomy. Phylogenetic trees are recognized as the background

information required for a huge number of inferences and explanations. But trees are simply a representation of which groups under examination form clades that do the real explanatory work. Essentially any question in evolutionary biology, or other branches of biology that make evolutionary assumptions, depends on history and so on clades.[20]

But in fact there is no single "common ancestry" relationship that grounds clade groupings. A standard definition of "clade" is that it is some species and all of its descendants. Yet it is not clear at all which groups are species. (Nor clear if there are any species.[21]) Further, some of the most popular views about species require that they are clades, and so at least those views cannot define "clade" in terms of species. For these reasons, it is now common to see clades defined directly in terms of groups of populations or organisms and their relationships.[22] But there are different, incompatible ways of understanding the history of populations and of organisms. Take these in turn.

Defenses of phylogenetic concepts of species often talk about trees of populations, to argue that all taxa (including species) should be monophyletic groups of populations.[23] That is, a clade should be some ancestral population and all of its descendants. This maneuver avoids talking about ancestral species, and avoids having delineation of clades depend on delineation of speciation events. But we then replace the avoided problem with the problem of delineating populations and population lineage splits. Velasco argues that lineage splits are context-dependent.[24] One rough

argument for this is that lineage splits represent a loss of cohesion between groups and the introduction of distinct evolutionary paths. However for certain kinds of traits a group may still be cohesive, while for others, the very same group may be broken up into independent trajectories. Only the context and associated conventions can determine which kinds of traits are of interest and so must help determine whether a lineage split has occurred.

The history of populations is naturally "loose" in a way that allows for some reticulation between groups. The very idea of migration dictates that it must be possible to have some gene flow between distinct populations without thereby collapsing them. How much reticulation is allowed is precisely what is at issue and what drives the point that lineage-splitting (and so cladehood) is context-dependent. Grant and Grant talk about distinct clades of Darwin's finches and place them on a phylogenetic tree, but later discuss hybridization between these groups.[25] There are many reasons to treat sister species of Darwin's finches as distinct clades. But whether the relevant lineages should be considered separate at all depends on context and convention.

This brings us to the history of organisms, because for some purposes, in some contexts, we want to be strict, and then it is important to think of clades as genealogically exclusive groups of organisms. That is: a group of organisms, all of which are more closely related to each other than to any organisms outside the group, with no exceptions such as hybrids. De Queiroz and Donoghue introduced this

concept of exclusivity to the taxonomic literature to separate it from monophyly in reticulating groups (such as organisms within a single species).[26] But there are different ways of understanding how organisms are related to one another. Baum and Shaw first carefully spelled out exclusivity in terms of genetic concordance, but Velasco defines it in terms of organismal parent-offspring relationships.[27] These two kinds of group are incompatible, with some biological projects concerned with one and different projects the other.[28]

Thus when we ask whether a group is genealogically exclusive, there is a suppressed variable that we might call *kind of genealogical exclusivity*. It can take (at least) the values "recency of organismal common ancestry" or "genetic concordance." But the biology alone does not determine which of these values the variable takes. So long as the available values are objectively incompatible as these two often are, any determination of whether a candidate group is genealogically exclusive is determination that we help with. This is because in such a typical case, our research interests, conventions, and so on, are involved in selecting among the available variable values. These contributions of ours must help select, if variable values are taken at all. Genealogical exclusivity is therefore conventional in our broad sense – determined by biology plus by us. When being a clade is being genealogically exclusive, we also help determine whether something is a clade.

We do not always want our understanding of common ancestry to be as strict as genealogical exclusivity, even though that exclusivity represents a kind of shared

ancestry that can ground many kinds of inferences. After all, a small number of hybrids between two different clades destroys either kind of genealogical exclusivity just described. And often we want to understand the distribution of some "broader level" feature such as biogeography, in which case it seems appropriate to think of the history of whole populations as determined by population lineage splits. But in these cases conventions help fix the variable value "distinct population lineage" in place of "being genealogically exclusive." And we saw that this fixed value itself has deeper suppressed variables, because population splits depend on contexts that have incompatible outcomes and which the biological facts alone do not choose between. So at multiple levels there is Indeterminacy Pluralism and conventionalism.

The general source of this is that different parts of a taxon have different histories. Which parts we care about varies across contexts. Our research interests help decide between the looser "population lineage" definition of clade or the more strict "genealogical exclusive group of organisms" idea. What is important to see is that on either of these readings, there are still further suppressed variables whose objective values would incompatibly dictate which things are population level lineages or which organisms are most closely related to each other. And the biological facts leave us with a plurality of possible values that lead to incompatible grouping of organisms into clades. Further details are needed for any determination of cladehood.

This is most obvious in extreme cases like Thermotogales. While much of the group's history remains uncertain, ribosomal RNA and other "core" operational genes give us

strong reason to believe that the Thermotogales are a bacterial group that share a "cellular" history with the bacteria Aquifilales; however, the majority of their genome indicates some other phylogenetic position – including many genes which are clearly of archaeal origins.[29] Context combined with various conventions helps determine whether Thermotogales is a clade of Bacteria, a clade of Archaea, or not a clade at all. While Thermotogales is among the most extreme cases we know, this kind of context dependence is unavoidable. There is then is no unique objective grouping of organisms into clades and so no uniquely correct tree of life.

6. Lumpers and Splitters

For splitters, those with a preference for finer-grained concepts, an objection now quickly comes to mind. Just as Dupré, Ereshefsky and Kitcher have split "species" into "genealogical species" and "interbreeding species" and others, objectivists can split the concept of "clade" into "population level clade" and "organism level clade". This may concede conventionalism about "clade", but relocates objectivism to the finer-grained concepts that the splitting of "clade" produces. Perhaps this splitter strategy can be applied to "functional cohesive unit" and "population" too.

Our above discussion indicates that the "clade" splitter's first division, between "population level clade" and "organism level clade," will only confront the further suppressed variables we have uncovered for each of these. Can the splitter then simply try and split yet again? Technically, yes. But consider the kinds of concepts that result:

- "population level clade" as clarified by a strict-but-not-too-strict "with <1 incoming migrant per generation per population"
- "organism level clade as defined by being an exclusive group due to being a clade on the plurality of the genome tree with respect to all genes and all organisms"

Putative objectivism about these concepts faces two problems. The first is that key components of these are also conventional. It is doubtful that biological facts alone fix what counts as a generation or a migrant, for example, much less what it takes to fix what counts as an organism or a population.

The second and more decisive problem is that these concepts, and the additional splitting that produced them, are theoretical dead ends. These concepts are not evolutionary grouping concepts at all. They are ad hoc constructions for the sole purpose of being objective and would play no role in biological theorizing. Being a clade is important. If we want to know what a clade is, we should focus on the role that the term "clade" plays in biological theory. But if we do this, it is clear that being a clade is tied up with many different kinds of processes, patterns, and methods of detection, and is fundamentally intertwined and interdefined with other "problem" concepts like lineage, species, and population. Being overly precise in defining "clade" robs the term of its power to play the large number of roles that it is expected to play.

This problem is even clearer for splitting "functional cohesive unit," which would give way to concepts such as:

- "functionally cohesive group with respect to trait T1 and pressure P1"
- "functionally cohesive group with respect to trait T2 and pressure P2,"

and...

If one of these applies to a group, it will typically apply to only that group—the one featuring T1 that is subject to P1, for example. Such concepts do not pick out kinds to which many member groups belong and over which theoretically interesting generalizations and predictions hold. Splitting "population" yields similarly vapid concepts:

- "population due to rate exclusivity with respect to survival interactions S1 and reproductive interactions R1"
- "population due to rate exclusivity with respect to survival interactions S2 and reproductive interactions R2"

The theoretical reasons for caring about the kinds of groups we have focused on in this paper are absent for those produced when avoiding our conventionalism by splitting. Better to not split and retain conventional concepts that are theoretically important. In other words, the utility of evolutionary group concepts depends on their flexibility in application. Empirical facts about the organization and diversity of the biological world dictate that our grouping concepts allow for flexibility in application in distinct contexts. It would be a mistake to conclude that this diversity requires instead a tremendous explosion in the number of evolutionary grouping concepts that we must use.

Lumpers will agree that splitting buys objectivity at too high a theoretical cost, and that we should retain instead the theoretically important concepts we have analyzed. But many lumpers will keep their agreement with us short, insisting that these unsplit concepts can objectively apply with theoretical significance across many contexts. For example, of course we would allow that in one context a group of populations is a functional cohesive unit if *all* their trait responses to *all* evolutionary pressures are more similar to each other than to any other populations. But the lumper argues that we also need to allow that a different group of populations can have the *same property* of being a functional cohesive unit with just 80% of their trait responses to all evolutionary pressures being more similar to each other than to any other population. In another context, 75% may suffice if this includes the right traits. In still others the difference may concern percentage of pressures instead of traits, but we still have a cohesive functional unit. We may need to recognize similar flexibility to being a population: in one context, exclusivity with respect to large sets or certain reproductive and survival interactions suffices, in another a smaller set of a different but especially salient mating interactions is enough. Perhaps a similar kind of consideration could apply to the different kinds of genealogical connections relevant to being a clade.

Effectively, these suggestions indicate that the unsplit concepts of "functional cohesive unit," Millstein's "population" and "clade" are all cluster concepts.[30] This is regardless of whether a particular group that one of these concepts picks out is a natural kind, an individual, or something else—in each case, the concepts are defined

by disjunctive clusters of conditions; no one of these conditions is necessary for application of the concept, but a variety of combinations of them is each sufficient.

There are two candidate ways to claim biological facts alone determine when such concepts apply. One is by putatively objective weighting schemes. Such a scheme hopes to tell us how much different factors matter in different cases. Notice that the weighting scheme that says that all factors matter equally across all cases in all contexts is still a weighting scheme (it is a strictly equitable one). The problem with any of these schemes is not that they are hopeless; they can be quite useful. The problem is for an objectivist reading of them. Thinking that biological facts alone determine which traits matter to what degree in what context *is* hopeless. Using the language of suppressed variables, our arguments in this paper have already implied these weightings are conventionally determined.

This fact inspires the second candidate way to claim that biological facts alone determine when cluster concepts apply in particular cases. The idea is that cluster grouping concepts cannot be defined in terms of the relative similarities or kinds of interactions between things that groups comprise. Granted, the disjunctive specifications of conditions above may be epistemic guides—they may help *indicate* whether a given concept applies to a group. But these conditions obtaining between things grouped are not what *makes* those groups evolutionary ones. Instead, the groups themselves have ontological priority as real units, whether as individuals (the populations of Millstein) or natural kinds (the species taxa of Boyd, *op. cit.*, and

Wilson, Barker and Brigandt, *op. cit.*). Reductive definitions of them are then bound to fail and the ontologies of evolutionary groups resist deeper specification.

This "groups first" view may seem a desperate way for the objectivist lumper to avoid the conventionalism of weighting schemes. But it is a difficult view to refute. We think it becomes more plausible when elements of it are retained in a conventionalist framework. In that framework, we appeal to biological facts, but also draw on our research interests, abilities, values, and so on, to conventionally determine which things are evolutionary individuals or kinds. Indeed, the least strained descriptions of biological practice accord with this. After conventional "group first" delineations pick out the groups, epistemically useful cluster specifications can be given and explain why these groups, rather than other possible groupings, fall under the relevant concept.

7. Conclusion and Broader Issue

Our chief conclusion is that evolutionary grouping concepts, such as those of clades, functional cohesive units, and populations, do not objectively delineate groups, but rather, when they do apply to some collection or organisms, they must do so with the help of our conventions. More specifically, facts about the context and our conventions, interests, and values combine with the biological facts to determine whether some collection of organisms is a group of a particular kind. The central reason for this concerns the suppressed variables built into these concepts. The concepts apply only when the variables take values. But the variables cannot take

values independently of context. And without our conventions helping to determine which values are taken, there is simply no fact of the matter about which collections are groups according to that concept. For just a flavor of how this conclusion is relevant not only to discussions of evolutionary groups, but also connects to a range of other issues, we finish with snapshots of broader implications.

The common assumption that the evolutionary groups we study form objectively determined branches on a single objective tree of all life is false. Prevailing taxonomic concepts each do not unambiguously divide sets of organisms into taxa when taking only objective biological facts as inputs. Instead, in different contexts, different groups are properly regarded as taxa.

Biological taxonomies have featured as case studies in broad "science and values" debates about whether and how our conventions and interests do or must attach to such taxonomies. Some objectivists concede that our conventions and interests help determine the *significance* of taxonomies, but we now see this is not a sufficient concession. Conventions and interests also help determine the *accuracy* of those taxonomies, because conventions and interests help fix the groups taxonomized.

Objectivity of evolutionary groups is often thought to contribute to or feature as a part of policy justifications. For instance, to help justify the conservation of certain caribou groups, the Committee on the Status of Endangered Wildlife in Canada (COSEWIC) can be understood as appealing to normative principles of the following

sort: if a described group of organisms objectively satisfies the definition of an evolutionary group concept (even a definition arbitrarily chosen among competitors), then the group is a unit of diversity that should be protected or conserved when endangered.[31] Even though the justification of this principle will probably be normative, it applies to any given group only if that group is objective. Policy makers, including COSEWIC, often suggest that the only or main obstacle to satisfying this objectivity condition on the application of the normative principle is a lack of empirical data (*ibid.*, p. 9). But our conclusion implies the obstacle is greater—that the objectivity condition cannot be met, nor the principle applied, no matter the state of empirical data.

---

[1] Philip Kitcher, *Science, Truth, and Democracy*. (New York, NY: Oxford University Press, 2001).

[2] Evidence for the age of the putative clade is discussed in A. Becker and G. Theissen, "The major clades of MADS-box genes and their role in the development and evolution of flowering plants," *Molecular Phylogenetics and Evolution* XXIX (2003): 464-89.

[3] Carrie Morjan and Loren Rieseberg, "How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles," *Molecular Ecology* XIII (2004): 1341-56.

[4] E.g., p. 77 in David A. Baum, "Species as ranked taxa", *Systematic Biology* LVIII (2009): 74-86.

[5] John Dupré, *The Disorder of Things: Metaphysical Foundation of the Disunity of Science* (Cambridge, MA: Harvard University Press, 1993); Marc Ereshefsky, "Eliminative Pluralism", *Philosophy of Science* LIX (1992): 671-690; Marc Ereshefsky, "Species Pluralism and Anti-Realism", *Philosophy of Science* LXV (1998): 103-120; Kitcher, *op. cit.*; and Philip Kitcher, "Species", *Philosophy of Science* LI (1984): 308-333.

[6] Ereshefsky, *op. cit.*; cf. Joseph LaPorte, Joseph, "Is there a single objective, evolutionary tree of life?," *The Journal of Philosophy*, CII (2005): 357-74.

[7] Lisa Gannett, "Making Populations: Bounding Genes in Space and in Time," *Philosophy of Science* LXX (2003): 989–1001.

[8] Kitcher (2001), *op. cit.*

[9] Kitcher, *ibid.*, Dupré, *op. cit.*, Ereshefsky *op. cit.*

[10] For syngameons see Alan R. Templeton, "The Meaning of Species and Speciation: A Genetic Perspective", in D. Otte and J. A. Endler, eds., *Speciation and Its Consequences* (Sunderland, MA: Sinauer, 1989), pp. 3-27. And for higher taxa see: Matthew J. Barker and Robert A. Wilson, "Cohesion, Gene Flow, and the Nature of Species," *The Journal of Philosophy* CVII (2010): 61-79; Marc Ereshefsky, "Species, Higher Taxa, and the Units of Evolution", *Philosophy of Science* LVIII (1991): 84-101.

[11] D. Brooks and D. McLennan, *The Nature of Diversity* (Chicago, IL: University of Chicago Press, 2002).

[12] Barker and Wilson, *op. cit.*; and Matthew J. Barker, "The Empirical Inadequacy of Species Cohesion by Gene Flow," *Philosophy of Science* LXXIV (2007): 654-665.

[13] At p. 1987 in T. Barbará, G. Martinelli, M. F. Fay, S. J. Mayo, and C. Lexer, "Population differentiation and species cohesion in two closely related plants adapted to neotropical high-altitude 'inselbergs', *Alcantarea imperialis* and *Alcantarea genicultata* (Bromeliaceae)," *Molecular Ecology* XVI (2007): 1981-92.

[14] Morjan and Rieseberg, *op. cit.*

[15] See pp. 74-75 in Baum, *op. cit.*

[16] Barker and Wilson, *op. cit.*

[17] Brent D. Mishler and Robert N. Brandon, "Individuality, Pluralism, and the Phylogenetic Species Concept," *Biology & Philosophy* II (1987): 397-414; and Barker and Wilson, *op. cit.*

[18] Roberta Millstein, "The Concepts of Population and Metapopulation in Evolutionary Biology and Ecology," in M. A. Bell, D. J. Futuyma, W. F. Eanes, and J. S. Levinton, eds., *Evolution Since Darwin: The First 150 Years* (Sunderland, MA: Sinauer, 2010), pp. 61-86.

[19] Peter Godfrey-Smith, *Darwinian Populations and Natural Selection*. (New York, NY: Oxford University Press, 2009).

[20] David A. Baum and Stacey D. Smith, *Tree Thinking: An Introduction to Phylogenetic Biology* (Greenwood Village, CO: Roberts and Company Publishers, 2012).

[21] Brent D. Mishler, "Getting Rid of Species?," in R. A. Wilson, ed., *Species: New Interdisciplinary Essays* (Cambridge, MA: MIT Press, 1999), pp. 307-315.

[22] Baum, *op. cit.*; Joel D. Velasco, "Species Concepts Should Not Conflict with Evolutionary History, but Often Do," *Studies in the History and Philosophy of Biological*

and Biomedical Sciences, XXXIX (2008): 407-414; Joel D. Velasco, "Species, Genes, and the Tree of Life", *British Journal for the Philosophy of Science* LXI (2010): 599-619.

[23] Velasco (2008), *op. cit.*

[24] Joel D. Velasco, "The Future of Systematics: Tree-Thinking Without the Tree," (submitted).

[25] Peter R. Grant, and B. Rosemary Grant, *How and Why Species Multiply: The Radiation of Darwin's Finches*. (Princeton, NJ: Princeton University Press, 2008).

[26] Kevin de Queiroz and Michael Donoghue, "Phylogenetic systematics or Nelson's version of Cladistics?," *Cladistics* VI (1990): 61-75.

[27] David A. Baum and Kerry L. Shaw, "Genealogical perspectives on the species Problem", in P. C. Hoch and A. G. Stephenson, eds., *Experimental and Molecular Approaches to Plant Biosystematics* (St. Louis, MO: Missouri Botanical Garden, 1995), pp. 289-303; Joel D. Velasco, "When monophyly is not enough: Exclusivity as the key to defining a phylogenetic species concept," *Biology & Philosophy* XXIV (2009): 473-86.

[28] Velasco (2010), *op. cit.*

[29] Olga Zhaxybayeva, Kristen S. Swithers, Pascal Lapierre, Gregory P. Fournier, Derek M. Bickhart, Robert T. DeBoy, Karen E. Nelson, Camilla L. Nesbø, W. Ford Doolittle, J. Peter Gogarten, and Kenneth M. Noll, "On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales", *PNAS* CVI (2009): 5865-70.

[30] See Richard Boyd, "Homeostasis, Species and Higher Taxa," in Robert A. Wilson, ed., *Species: New Interdisciplinary Essays* (Cambridge, MA: MIT Press, 1999), pp. 141-185; and Robert A. Wilson, Matthew J. Barker, and Ingo Brigandt, "When Traditional

Essentialism Fails: Biological Natural Kinds," *Philosophical Topics* XXXV (2007): 189-216.

[31] D. Thomas and D. Gray, "Update COSEWIC Status Report on the Woodland Caribou *Rangifer tarandus caribou* in Canada," *Committee on the Status of Endangered Wildlife in Canada* (Ottawa, 2002).

Philosophy of Science

Time and Fitness in Evolutionary Transitions in Individuality

--Manuscript Draft--

Manuscript Number:	11556
Full Title:	Time and Fitness in Evolutionary Transitions in Individuality
Article Type:	PSA 2012 Contributed Paper
Keywords:	Evolutionary Transitions in Individuality; ETI; Fitness; Time; Levels of Selection
Corresponding Author:	Pierrick Bourrat University of Sydney Sydney, NSW AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Sydney
Corresponding Author's Secondary Institution:	
First Author:	Pierrick Bourrat
First Author Secondary Information:	
Order of Authors:	Pierrick Bourrat
Order of Authors Secondary Information:	
Abstract:	It is striking that the concept of fitness although fundamental in evolutionary theory, still remains ambiguous. I argue here that time, although usually neglected, is an important parameter in regards to the concept of fitness. I will show some of the benefits of taking it seriously using the example of recent debates over evolutionary transitions in individuality. I start from Okasha's assertion that once an evolutionary transition in individuality is completed an ontologically new level of selection emerges from lower levels of organization. I argue that Okasha's claim to have identified two ontologically distinct levels of selection is an artifact created by an undeserved comparison between the fitness of the collective level and the fitness of its constituents. Once fitness is assessed over the same period of time at the two levels of organization it becomes clear that only one, unique process of selection is acting upon both levels.

Manuscript

Time and Fitness in Evolutionary Transitions in Individuality

Abstract

It is striking that the concept of fitness although fundamental in evolutionary theory, still remains ambiguous. I argue here that time, although usually neglected, is an important parameter in regards to the concept of fitness. I will show some of the benefits of taking it seriously using the example of recent debates over evolutionary transitions in individuality. I start from Okasha *'s* assertion that once an evolutionary transition in individuality is completed an ontologically new level of selection emerges from lower levels of organization. I argue that Okasha *'s* claim to have identified two ontologically distinct levels of selection is an artifact created by an undeserved comparison between the fitness of the collective level and the fitness of its constituents. Once fitness is assessed over the same period of time at the two levels of organization it becomes clear that only one, unique process of selection is acting upon both levels.

1. Introduction

> Yes, fitness is the central concept of evolutionary biology, but it is an elusive concept. Almost everyone who looks at it seriously comes out in a different place.
>
> Leigh Van Valen 1989,2-3

It is striking that the concept of fitness, although fundamental in Darwinian Theory, is not yet unified, and after more than 150 years still remains ambiguous. Is fitness an ecological descriptor or a mathematical predictor? Do species have a fitness, and if they do, how shall we measure it? Should fitness be measured over short or long periods of time? All these questions are still without clear answers. In this article, I have two aims. First, I will clarify the concept of fitness by arguing that time is an important parameter of this concept. Discussions over the concept of fitness are numerous and I will not be able to cover them all. Rather, I will concentrate on particular benefits that pairing the concept of fitness with time can bring to one contemporary discussion over the levels of selection, namely evolutionary transition in individuality (ETI). I will demonstrate, and this will be my second aim, that the model of ETIs developed by Okasha (2006), relying on Michod and colleagues' work, faces a serious problem. This problem, I will argue, comes precisely from the fact that in his model Okasha does not sufficiently take time into account when measuring fitness at different levels of organization.

ETIs are events in the course of evolution that lead to the formation of new higher level individuals due to the cooperation of two or more individuals at a lower level of organization (Michod 2011). One example of ETI is the transition from uni- to multicellular organisms. A number of other ETIs have been proposed, among them the transitions from prokaryote to eukaryote cells, from unicellular to multicellular organisms, and from multicellular organisms to integrated colonies such as colonies of ants or honeybees. One of the most accomplished models of ETI is the one suggested by Michod and colleagues. In a number of articles and books (Michod 1999, 2005; Michod, Nedelcu, & Roze 2003; Michod, Viossat, Solari, Hurand, & Nedelcu 2006) they propose a number of conditions for ETIs to occur. Okasha (2006, 2011) recently set Michod and colleagues' work in the framework of Multilevel Level Selection 1

(MLS1)/ Multi Level Selection 2 (MLS2), which was initially developed by Damuth and Heisler (1988).

Okasha's and Michod's models of ETIs are committed to a concept of fitness which is measured by the ability of a given entity to survive and reproduce in its environment. According to this definition, the higher the survival and reproductive rate of this entity, the higher its fitness is. Although such definition is somewhat restrictive and does not cover the whole range of possibilities which can be embraced by the concept of fitness, I will accept it as common ground for the development of my arguments, which will run as follows. In Section 2, I will briefly review Michod and colleagues' as well as Okasha's models of ETIs in regards to fitness. I will present two specific claims defended by both authors: (1) that during the last stage of an ETI, once a division of labor is in place, the fitness of the components constituting the newly emerged individual reaches zero; (2) that there are two fundamentally distinct processes of selection, namely multilevel selection 1 (MSL1) and multilevel selection 2 (MLS2), occurring alternately at the different stages of an ETI. Claims (1) and (2) are slightly different versions of what is called the export-of-fitness view on ETI. Claim (1) has been recently criticized by Godfrey-Smith (2011, 77-78) for its metaphorical nature. Although this criticism deserves a more thorough examination, that will not be done in this article. In Section 3, I turn to claim (2) and demonstrate that if fitness is assessed over the same period of time at the collective level and at the level of its constituents, then there is commensurability between these two models of selection. For that reason, they cannot represent two ontologically distinct processes of selection, but are ways to describe the same process from the perspective of two spatial and two temporal scales. However, I do not deny the epistemological value of describing ETIs within the MLS1/MLS2 framework and will examine the reasons for this.

2. Michod and Okasha on evolutionary transitions in individuality

Michod and colleagues propose the following model of ETI. For new individuals at a higher level ("collective" level) to emerge from a lower level ("particle level")[1], e.g. for multicellular organisms to emerge from unicellular organisms, two things must happen. First, conflicts between members of the collective need to be eliminated. Conflicts can be resolved in different ways such as for instance policing mechanisms and developmental bottlenecks, to name two of them. They both promote genetic homogeneity and consequently reduce competition between the different members of a group. However, even if genetic homogeneity is reached between the different members of the same group, this will not necessarily lead to the emergence of a higher individual. For an ETI to take place, Michod and colleagues propose that there must be a division of labor between germ and soma (or its equivalent in ETIs other than from uni- to multicellular organisms), since without it, the collective fitness will be proportional to the average particle fitness. As such, the collective will not be an individual with its own fitness (Michod 2005, 569); its fitness will merely be a cross level by-product of its particles' fitness.

Claim 1

As I noted earlier, Michod and colleagues define the fitness of an entity (whether particle or collective) as the product of its viability and fecundity, which is often done in life-history models. In the cases of transition from unicellular to multicellular organisms with full separation

[1] The distinction between particle and collective comes from Okasha (2006, 4)

of germ and soma, if a cell does not specialize, it will invest its resources in both the viability and fecundity components of fitness. As a result its fitness will be positive. However, Michod (2005, 2011) and Okasha (2009) both generalize this argument over other ETIs and propose that:

(1) If a particle invests everything in the somatic (or germ) function (or its equivalent) of the future collective individual, it will have a fitness equal to 0, since although its viability (or fecundity) component of fitness will be positive, its fecundity (or viability) component and consequently the product of viability and fecundity will be nil.

However, when the two types of particles combine their investment in both components of fitness (one investing everything in the soma and the other everything in the germ function) a new collective individual emerges with its own fitness. This reasoning leads Michod and colleagues to claim that during an ETI transfer of fitness from the particle to the collective level.

Claim 2

Okasha (2006) and Michod (2005, 2011), mostly relying on Okasha's analysis, both link this work to the two concepts of multilevel selection distinguished by Damuth and Heisler (1988), namely MLS1 and MLS2. In the MLS1 framework, the focal unit of selection is the particle. For that reason fitness is expressed in a number of particles produced. For example, a group of particles will have a higher fitness than another if ceteris paribus it produces more particles. In MLS1, the fitness of the collective is merely a "by-product" of the different fitnesses of the particles composing this collective. In the MLS2 framework, the focal units of selection are both the particle and the collective. Fitnesses of the collective and of the particle are measured in different units. The fitness of a collective is expressed in number of new groups it produces

independently of the number of particles each group is composed of, while the fitness of a particle is simply expressed in number of particles it produces. During an ETI, Okasha (2006, 237-238) argues, there are three stages for which MLS1 and MLS2 are alternately more relevant to describe the selection process and propose that:

(2) MLS1 and MLS2 are two distinct causal processes of selection as opposed to two conventional ways of expressing selection (2006, 59; 2011, 243). During an ETI, they represent a transition in processes of selection. Not only MLS1 and MSL2 are alternately more relevant at the different stages of an ETI, they are alternately the only way to describe accurately the process of selection.

In the first stage of an ETI, the particles of the future collective start to aggregate and cooperate. The fitness of this newly formed collective is merely the average of the particles' fitness, hence MLS1 is the relevant type of selection occurring. During the second stage, the fitness of the collective is not defined in terms of the particles any more, but is proportional to the average fitness of the particles. At that stage, although MLS2 framework can be applied, so can MLS1. There is a "grey area between MLS1 and MSL2", in Okasha's words (2006, 237). However, the collective lacks individuality, since its fitness is a cross-level byproduct of the particles' fitness. During the third stage, when the transition is complete, the fitness of the collective cannot be expressed as the average fitness of the particles any more. The collective is now an individual on

its own and its fitness is not proportional to the fitness of the particles; both fitnesses are now incommensurable[2].

3. When time makes a difference

Where does the incommensurability between particle and collective fitnesses come from? To this question there is no clear answer and it is not clear how there could be one even in principle. It is in fact hard to imagine that collectives could exhibit variations in fitness, without their constitutive parts exhibiting a form of variation with consequences on their own fitness. Yet Okasha believes that such scenarios exist (Okasha 2006, 106) and that they materialize when MLS2 is the framework of choice, for MLS2 framework, he claims, fits two causally distinct processes of natural selection happening in nature (Okasha 2006, 59; 2011, 243). Recall that in MLS2 framework, the fitness of the collective can be defined as a quantity "that bears no necessary relation to average particle fitnesses alone" (2006, 136, my emphasis). Yet, in the same sentence Okasha surprisingly asserts that "it is impossible that the resulting evolutionary change could be expressed in terms of particle fitnesses alone," Okasha (2006, 136, my emphasis). Beyond, the fact that the consequence does not follow from the premise (Okasha should have used "sometimes impossible" instead of "impossible"), I propose one important reason why we should doubt this claim in any case. I will not argue here either against the MLS2 framework itself since it is obviously mathematically true. Rather, I will argue against the claim

[2] Michod and colleagues use the word 'decoupling' to refer to this phenomenon. By decoupling they mean that the fitness at the collective level becomes expressed in a different currency than fitness at the particle level and that it is not translatable into fitness at that level

that there is incommensurability between the particle and collective fitnesses in any real cases of evolution by natural selection. The reason I will give is based on purely methodological grounds linked to time, fitness and levels of organization and will be illustrated with one of Okasha's own example of MLS2.

In chapter 7 Okasha (2006) deals with species selection, the paradigmatic case of MLS2 in the literature on the subject, and embraces Vrba's 'acid test' (1989, 155) to detect true species selection (and more generally MLS2) from mere by-products of selection at lower levels, as in MLS1. Vrba proposes that there is true species selection when the outcome of selection at the species level cannot be explained from the perspective of the organism. One stringent way to know when this happens is to seek different directions of selection at the different levels of organization. For instance, species selection, if truly independent, could in principle counteract selection at the organism level. Vrba's test will however be inconclusive when both selection processes push in the same direction, but the most reasonable attitude to adopt in such case will be to consider that selection only really occurs at the lower level, unless one would be able to display that the force at the species level has different value from the force at the organism level. Okasha claims that one example of true species selection satisfying Vrba's test is involved in the evolution/maintenance of sexual reproduction. He asserts that asexuality is advantageous at the organism level, because of the two-fold cost of producing males (Maynard Smith 1978), but that sexuality is advantageous at the species level because it allows faster evolutionary responses to rapid changes in environmental conditions. According to this reasoning, sexual lineages would be selected via species selection as a distinct process of natural selection different from selection at the organism level which favors asexual organisms.

One fundamental principle of the scientific method in experiments is to change only one variable at a time while the other are kept unchanged or controlled. To reach this goal, if one is interested in measuring the influence of X (a drug, for instance) on a population P, the experimenter will need to control the effect of X on P with another population (let us call it P' or Control) which was not administered X but which is as similar to P as possible in all other respects. Hence, if a difference is observed between the two populations, it will only be attributable to X because no other variable will be different. However, if P and P' are not strictly identical in all respects but X, then any observed difference could be attributable to X or any of the other different variable between the two populations and which could have the same effect than X. Such variables are called confounding variables. How is that relevant to our problem of species selection and Vrba's test? Vrba's test is not a scientific experiment per se, but it shares with them the necessity to be controlled. Unless all the variables relevant to selection are strictly identical at both levels in the test, the detection of a different direction of selection at those levels could be attributed either to a different process of selection at each level or to any other variable with different values at each level and with some relevance to selection. Just like any scientific experiment, Vrba's test requires that only one variable at a time is changed while all the other are kept unchanged.

We noted earlier that Okasha claims that the evolution/maintenance of sexual reproduction is a true case of species selection. He justifies this assertion using Vrba's test. Because, he argues, the test shows that selection pushes in two opposite directions (i.e. sexuality at the species level and asexuality at the organism level), a process of selection ontologically different from the process of selection at the organism level, must exist at the species level. But does Okasha's comparison eliminate all possible confounding variables, which would render his conclusion

spurious? In other words, is selection at the organism level assessed in the exact same way at the species level? The answer to this question is that it is not; a confounding variable does exist.

To detect this confounding variable, let us consider two types of organisms, one asexual and one sexual, under the same selection pressures. To reproduce, sexual organisms spend energy both to look for a partner and to produce gametes during meiosis, while only half of their genes will be represented at the next generation. On the contrary, asexual organisms will be able to reproduce genetically identical offspring, without any cost from meiosis or courtship and mating. Hence if the two types of organisms are in competition, the asexual ones should quickly out-compete the sexual ones, because of the supplementary costs associated to sexual reproduction. At that point, it is thus extremely tempting to claim that the fitness of an asexual organism is higher than the fitness of a sexual organism. But, if formulated as such, this claim would be incomplete and would have to be relativized over a period of time (e.g. one generation).

Why is that? First, because the fitness of an organism cannot be directly measured as, for instance, the mass of an object can be; measures of fitness are only proxies for fitness. Second, because different proxies for fitness can lead to different answers. Hence, the information about the way fitness is measured is always relevant. In fact, the reproductive output after one generation of an organism represents only one proxy for its fitness among an infinite number. There is no "best" period of time over which one can measure fitness. This type of problems leads Beatty & Finsen (1989) and Sober (2002) to propose a distinction between short-term and long-term fitness. In most cases the short-term reproductive output of an entity is a good proxy for fitness to grasp the evolutionary dynamics of interest. But at other times it might be insufficient, and we will need a proxy measuring the reproductive output over a longer period of time. One famous case, proposed by Fisher (1930) on sex ratio, makes the reproductive output

two generations ahead a much better proxy for fitness than one generation. More generally, proxies of fitness over long periods of time should be preferred if one is interested in evolutionary problems involving changes in the environment, as it is the case with the evolution and maintenance of sexual reproduction. This is because long term environmental changes and their consequences on selection pressures will be invisible to a proxy for fitness based on the short term reproductive output. Yet, many evolutionary problems do not involve such changes and measuring fitness as the reproductive output over one generation is fine because the environment usually does not change or changes very little over one generation. This is the case for instance if one wants to know what phenotype is optimal in a constant environment.

The confounding variable in Okasha's comparison becomes now obvious. It is the time over which fitness is assessed, which is itself a proxy for environmental changes. At the organism level, fitness is usually measured as the reproductive output after one organism's generation. At the species level, fitness is measured as the rate of extinction or speciation over much longer periods of time, sometimes many millions of years. But commensurability necessarily exists between fitness of species and fitness of organisms. Speciation and extinction events are ultimately composed of the deaths, survivals and reproductions of organisms over many generations, since the former events supervene on the latter ones. Thus, when Okasha applies Vrba's test over the maintenance/evolution of sex, he compares the fitness of organisms over one generation at the organism level with the fitness of organisms over a much higher number of generations[3]. Performed as such, Vrba's test remains inconclusive. Indeed, the difference observed could be either due to two processes of selection pushing in two opposite direction or to

[3] In virtue of the supervenience of speciation and extinction events at the species level on death, survival and reproduction events at the organism level

two measures of one and the same process of selection over two different periods of time, pushing in one direction over the short term and in the other over the long term. In the rest of the article, I defend the latter possibility.

To see why, let us now perform Vrba's test while controlling the period of time over which fitness is measured. Controlling time could be done in two ways: (a) by measuring fitnesses at both the species level and the organism level over one organism generation and compare them over this period of time; (b) by measuring the two fitnesses over the period time that would normally be used to measure species' fitness, that is, a period long enough to detect events of speciation or extinction. Both alternatives seem to be doomed in practice, since we are neither able to measure the fitness of species over short periods of time, nor able to measure the fitness of organisms over periods of time longer than a few generations. But if we were able to do so, we would certainly find that ceteribis paribus asexual organisms and asexual species have a higher short-term fitness as measured by (a) than sexual organisms and sexual species, but have a lower long-term fitness as measured by (b). The reason for that is not mysterious. Asexual organisms and asexual species on average do better when the conditions are stable (as it is usually the case over one generation) while sexual organisms and sexual species do better when new environmental conditions arise (which certainly occur over several millions of years). In other words, both selection at the organism and the species level would go in the same direction once the test is controlled for the period of time over which fitness is measured.

Thus, Okasha's claim that the evolution/maintenance of sexual reproduction occurs as a result of species selection is inexact. If we follow his reasoning using time as a constant over fitness (itself as a proxy for the stability of the environment), we predict no difference between a measure of selection made at the level of the organism and another one made at the level of the

species. The most natural implication is that these different measures represent one and the same process of natural selection, but expressed in different terms and over different periods of time.

There is no logical barrier to extending this argument to all the other cases for which MLS2 has been the framework of choice. In each case, if fitness could be determined over the same period of time or in the same constant environment at each level, what seems to be ontologically different levels of selection could in principle be unified under one and the same process. Does it mean that MLS2 framework should be abandoned and always replaced by MLS1? I claim that it should not, unless one has the full availability, at any point in time, of the selection pressures on the particles under consideration. I can only see multilevel models as satisfying these criteria. In any real case, the complete list of selection pressures will be most of the time unknown or they will be constantly changing (e.g. frequency dependent selection) making thus the particle fitnesses extremely complex to determine over long period of time. When both particle and collective fitnesses are available, and that the question at stake is about the collective, I propose that the MLS2 framework should be privileged. There are two further reasons for this choice. First, the complex task of measuring fitness of all the particles within a collective (with all the non-linear relations it implies) and over many particles' generations will often materialize at the collective level into a single and easily measurable parameter: the collective's reproductive output. Second, keeping fitness of the particles and fitness of the collectives independent, as it is done in MLS2 framework, can bring different, yet relevant, information about the selection since they are measured over different periods of time.

After these general consideration on MLS2 what does the distinction MLS1/MLS2 become in the context of ETIs and especially during their last stage? Would it be, in principle, possible, at the last stage, to describe the fitness of a collective in terms of the fitnesses of its particles,

contra Okasha? Following the reasoning I used in the case of the evolution/maintenance of sex, as in any case of MLS2, I see nothing that would prevent it. During an ETI, if the fitness of the particles seems incommensurable with the fitness of the collective, it is most probably due to the fact that, during the last stage, both fitnesses are not measured over the same period of time anymore and that the interactions between particles become so complex that tracking back their fitness over longer periods of time than one or two generation appears in practice impossible. What becomes decoupled in the two levels is not fitness per se but generations or life cycles. Because Michod's proxy for fitness depends on reproductive output after one generation, if "one generation" does not mean the same thing at the particle and the collective level, it is not surprising that collective and particle fitnesses seem decoupled from each other. But this is an artifact created by the measure. That does not mean that MLS2 represents an ontologically distinct process of selection from MLS1. Rather, it suggests that MLS2 is very useful means to carve one single processes of natural selection both in time and space and becomes especially useful once an ETI is completed. This echoes a recent criticism made by Waters (2011) about Okasha's fundamentalism over the distinction between MLS1/MLS2 in which he claimed that MLS1 and MSL2 frameworks were conventional rather than fundamental. Okasha (Okasha 2011, 243) held his ground, restating that they were fundamental. I have provided evidence here that they clearly were conventional and it became apparent once measures of fitness were controlled over time.

4. Conclusion

I have demonstrated that time is an extremely important parameter to take into account in regards to the concept of fitness. I argued for its relevance in ETIs and, more generally, in the levels of selection debate. I used the evolution/maintenance of sexual reproduction as a case study to establish that if different proxies of fitness reflecting different time scales are used at the organism and species levels, this will have the consequence of measuring selection pressures over two different time scales. This can lead one to confound the existence of one unique process of selection over two different periods of time with two ontologically distinct processes of selection, one for each level. I applied the same reasoning to ETIs and argued that they were not transitions in processes of selection, but rather events for which MLS1 and MLS2 were, although ultimately formally equivalent, alternately more relevant. The claim that distinction between collective selection and particle selection is conventional is not new (e.g. : Dugatkin & Reeve 1994; Sterelny 1996) and Kerr & Godfrey-Smith (2002) have formalized this equivalence. Yet, as Okasha (2006, 136) rightly points out, this formalism has been made solely in the context of MLS1. Taking time as an important variable in measures of fitness represents one important step towards a formalism in which events of selection normally described under the MLS2 framework, such as the last stage of ETIs, could also be, described under the MLS1 framework.

References

Beatty, J., & Finsen, S. (1989). Rethinking the propensity interpretation: A peek inside Pandora's box. In M. Ruse (Ed.), What the Philosophy of Biology Is: Essays Dedicated to David Hull. Dordrecht: Kluwer Publishers.

Damuth, J., & Heisler, I. L. (1988). Alternative formulations of multilevel selection. Biology and Philosophy, 3(4), 407-430.

Dugatkin, L. A., & Reeve, H. K. (1994). Behavioral ecology and levels of selection: dissolving the group selection controversy. Advances in the Study of Behavior, 23, 101-133.

Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Oxford: Clarendon Press.

Godfrey-Smith, P. (2011). Darwinian populations and transitions in individuality. In B. Calcott & K. Sterelny (Eds.), The Major Transitions in Evolution Revisited. Cambridge, MA: MIT Press.
Kerr, B., & Godfrey-Smith, P. (2002). Individualist and multi-level perspectives on selection in structured populations. Biology and Philosophy, 17(4), 477-517.
Maynard Smith, J. (1978). The Evolution of Sex: Cambridge University Press.
Michod, R. E. (1999). Darwinian dynamics. Princeton: Princeton University Press.
Michod, R. E. (2005). On the transfer of fitness from the cell to the multicellular organism. Biology and Philosophy, 20(5), 967-987.
Michod, R. E. (2011). Sex and multicellularity as evolutionary transitions in individuality. In B. Calcott & K. Sterelny (Eds.), The Major Transitions in Evolution Revisited. Cambridge, MA: MIT press.
Michod, R. E., Nedelcu, A. M., & Roze, D. (2003). Cooperation and conflict in the evolution of individuality: IV. Conflict mediation and evolvability in Volvox carteri. BioSystems, 69(2-3), 95-114.
Michod, R. E., Viossat, Y., Solari, C. A., Hurand, M., & Nedelcu, A. M. (2006). Life-history evolution and the origin of multicellularity. Journal of theoretical Biology, 239(2), 257-272.
Okasha, S. (2006). Evolution and the Levels of Selection: Oxford University Press, USA.
Okasha, S. (2009). Individuals, groups, fitness and utility: multi-level selection meets social choice theory. Biology and Philosophy, 24(5), 561-584.
Okasha, S. (2011). Reply to Sober and Waters. Philosophy and Phenomenological Research, 82(1), 241-248.
Sober, E. (2002). The two faces of fitness Thinking about evolution: historical, philosophical, and political perspectives. Cambridge: Cambridge University Press.
Sterelny, K. (1996). The return of the group. Philosophy of Science, 63(4), 562-584.
Van Valen, L. M. (1989). Three paradigms of evolution. Evolutionary theory, 9(1), 1-17.
Vbra, E. S. (1989). Levels of selection and sorting with special reference to the species level. Oxford Surveys of Evolutionary Biology, 6, 111-168.
Waters, K. C. (2011). Okasha's Unintended Argument for Toolbox Theorizing. Philosophy and Phenomenological Research, 82(1), 232-240.

Presentism as an empirical hypothesis

Abstract Within philosophy of physics it is broadly accepted that presentism as an empirical hypothesis has been falsified by the development of special relativity. In this paper, I identify and reject an assumption common to both presentists and advocates of the block universe, and then offer an alternative version of presentism that does not begin from spatiotemporal structure, which is an empirical hypothesis, and which has yet to be falsified. I fear that labelling it "presentism" dooms the view, but I don't know what else to call it.

1. Introduction

Here are two premises:

(P1) All and only things that exist *now* are real.

(P2) Special relativity is a complete account of spatiotemporal structure.

The first is a version of so-called "presentism". It says that what is *real* is what exists *right now*: this is what there is. Things in the past aren't real (they don't exist any more); things in the future aren't real (they don't exist yet). What there is, is what is *present*. "Now" bears a lot of ontological weight. The second says that there are no good reasons for adding anything to the account of space and time found in

Einstein's special theory of relativity: it's complete. The problem for people attracted to presentism is that (P2) seems to be incompatible with (P1), for reasons that I'll come to below. The dominant view amongst philosophers of physics is that we should therefore reject (P1) and adopt the so-called "block universe", a four-dimensional structure where everything that has ever existed, and ever will exist, is all equally real. There is no "now" of any ontological significance, there's just the whole four-dimensional shebang. That's what there is. What Einstein showed in developing special relativity is that, even if there is such a thing as "the present", we have no empirical access to it. If there is any evidence for presentism, then it does not come from empirical experience.There is nothing in empirical experience that supports this concept; there is no empirically well-grounded concept of "the present". Presentism, if treated as the hypothesis that there is such a concept, is false.

For philosophers of physics such as myself, this "block universe" is the default position, the "well yeah, of course" point of view, according to which anyone who's a presentist hasn't learned the hard-won lessons from physics properly. End of story.

However, in this paper I argue for an alternative empirical approach to "the present", equally well founded in physics as Einstein's treatment of simultaneity, according to which the dispute between presentists and block universe people remains an open empirical question, not decided by special relativity. In other words, even if you endorse (P2), there's a form of presentism that remains a live option. I was sufficiently surprised to find myself reaching this conclusion that I decided I should write it down.

2. Why presentism is false

Let's go back to the two premises above, (P1) and (P2), and remind ourselves why they appear not to work very well together. According to Howard Stein (in a 1968 paper responding to Putnam's 1967

paper on special relativity), adopting both (P1) and (P2) leads to "the interesting result that special relativity implies a peculiarly extreme (but pluralistic) form of solipsism." The reason for this is familiar from the literature on space and time. Given an event e1 in spacetime, e1 is "now" relative to itself, but there is nothing within the structure of special relativistic spacetime that determines which events spatiotemporally distant from e1 are *also* "now" relative to e1. There's no preferred way to join the dots and say these two events are both "now". You can conclude this *directly* from conventionality of simultaneity, in which case any "joining of the dots" in planes of simultaneity is an addition, going beyond the content of special relativity. Or you can get there via relativity of simultaneity: adopt the Einstein synchrony convention, note that different planes of simultaneity make different determinations of which events are "now" relative to e1 and which are not, note that picking *one* of these – a *preferred* plane of simultaneity – goes beyond the content of special relativity, and so conclude that no other events are determinately "now" with respect to e1. Either way, the conclusion is that there's no preferred way to join the dots.

If we focus our attention on (P2), and ignore (P1), this argument typically leads to endorsing the block universe. What special relativity gives us is just the entire set of events, arranged in a four-dimensional block.

But if we want to have (P1) as well as (P2), we get a different conclusion. If no other events are determinately "now" with respect to e1, then, by (P1), no other events are determinately "real" with respect to e1. Add to that the claim that everything that is real must be determinately so, and we get our conclusion: nothing is real with respect to e1 except e1 itself (hence the extreme solipsism), although each event is real with respect to itself (hence the pluralism). This is Stein's pluralistic extreme solipsism.

"Pluralistic extreme solipsism" is what you get if you hang on to presentism as expressed in (P1), and to (P2). So far as I know, the view hasn't attracted many adherents, and I suppose that's not very

surprising. Instead, the standard moves in the interpretation of special relativity reject either (P1) or (P2). On the one hand, there are those who suggest that "taking our experience of time seriously" requires us to reject, or supplement, special relativity. So we accept (P1), and in an effort to avoid the slide towards extreme solipsism, we reject (P2), perhaps trying to stay as close as possible to special relativity but adding a preferred foliation so that we get a unique, global "now". On the other hand there are those who suggest that "taking special relativity seriously" requires us to give up presentism. We endorse (P2) and straightfowardly reject -- throw out -- (P1). We adopt the four-dimensional "block universe". And we characterize our presentist opponents as intellectual cowards, clinging to their unfounded pre-critical intuitions in the face of overwhelming evidence from the conceptual developments wrought by science. There are lots of variations, lots of ways of trying to finesse things, but those are the basic moves.

I am in agreement with the proponents of the "block universe" in this dispute, as I have set it up. I don't think there *is* some "every day" concept of time that we can make use of philosophically and that's independent of the scientific concept. Science starts from everyday experience and investigates those very concepts, clarifying and changing them along the way. DiSalle makes this point very beautifully in his book *Understanding spacetime*. The book is about the engagement of physics with our concepts of space and time: the way that developments in physics have brought about developments in those very concepts, and how there's no "other" concept of time that's independent of these developments and somehow left standing untouched by them.

Presentism as expressed by (P1), and in appropriately similar versions, as an empirical hypothesis is false.

3. A shared assumption

Below, I argue for an alternative formulation of presentism, as an empirical hypothesis, such that the considerations of simultaneity just discussed in section 2 do not lead to the conclusion that presentism is false. I will *not* do this on the basis of our "experience of time". Rather, I will make the argument from within physics itself. I endorse (P2): special relativity gives us a complete account of spatiotemporal structure. I will not add a preferred foliation or anything like that; I think that's entirely misguided. Instead what we need to do is investigate the conceptual development that has taken place, with philosophy and physics taken together, working hand-in-hand, and when we do that we see that there *is* a presentist alternative available, and one which will *not* be "pluralistic extreme solipsism".

The reason why I reject "pluralistic extreme solipsism" is not just distaste, but because there is an assumption in the interpretation of special relativity that's common to both the "block universe" and the presentist positions characterized above, which I think the presentist should reject. The reason *why* "pluralistic extreme solipsism" follows from adopting both (P1) and (P2) is that space-time is being used by both of the block universe people and the presentists as a principle of ontological unity. This is the "shared assumption" to be rejected, and it is a familiar claim according to which space and time provide the ontological framework within which everything that is material exists. Or, to put it another way, spacetime is the ground of the unity of the world: what makes this material universe *one* universe is the shared space and time framework within which the matter is located.

For those of us interested in modern science, this approach to the unity of the world has a venerable pedigree. In Newton's physics, space and time can be understood as playing just such a role.

With this assumption explicitly on the table, let us revisit (P1). (P1) attempts to ground the unity of what exists, of what is real, in simultaneity. All and only things that exist *now* are real. If the "now" of a given real thing extends to other things, then those other things are also real. The unity of the real is grounded in their simultaneity. That's okay, if there's absolute simultaneity. But if "now" is *not* spatially extended, then what is real – given P(1) – is not spatially extended either. As discussed above, special

relativity *doesn't* underwrite a spatial extension of "now" via absolute simultaneity and so, given a commitment to special relativity as asserted by (P2), we arrive at pluralistic extreme solipsism.

What's gone wrong is that we bought into the proposal that spacetime is the ontological ground of the unity of what there is. My argument begins from the observation that we don't *have* to use spacetime to play this role. In what follows, I reject the shared assumption and develop an alternative formulation of presentism accordingly.

4. An alternative ontological principle of unity

Something that I won't argue for in this paper, but will simply state, is that Newton's physics contains alternatives to space and time as the ontological ground of the unity of what there is. One of these grounds unity in the three laws of motion. According to this approach, we begin not with space and time, but with the laws, which are viewed as providing a principle of unity. The proposal is that we don't begin with the physical entities, as given unities, and then ask about the laws that they satisfy. Rather, the laws themselves play a constitutive role in constituting the very entities that are their subject-matter, and in constituting them as genuine unities. Thus, to be a physical thing, be it simple or composite, *is* (in part) to conserve total quantity of motion (when isolated from other physical things), this being required by Newton's laws. Moreover, the laws also play a constitutive role with respect to the parts of a composite physical entity. Consider, for example, a collision between two billiard balls: not only is the total quantity of motion conserved (so the composite is a genuine unity), but that quantity is redistributed determinately, and from the law-constitutive perpective this is what makes the billiard balls themselves genuine unities throughout the process. The unity of the whole, and the unity of each part, is grounded in the laws. I cannot argue for this view here, for that would be another paper in itself. I make use of the view in what follows, in discussing presentism, and in so doing will further elaborate it

somewhat. One important aspect of the view will be its treatment of change. Changes in the state of a component are determined by the laws, so the laws provide an account of what it is for a genuine unity (the part) to undergo change whilst remaining the very same thing. Saying the same thing another way: the laws provide an account of what it is for a unity to persist through change: that is, to retain its numerical identity whilst not its qualitative identity. It does this without appeal to either essential properties or to haecceities. It offers us an alternative and one which, I am going to argue, favors a version presentism that is an empirical claim, but one which is distinct from that explored by Einstein in his 1905 paper on special relativity. For now, I am simply making a claim: there is in Newton's physics an alternative option for a principle of unity, based on the laws, and this should be on the table as something that we discuss and evaluate.

5. Change

What is it for an object to persist through change? The *prima facie* puzzle here is as old as it is familiar. How can a thing – by which we mean a genuine unity – remain the very same thing and yet undergo change? In particular, if F and G are inconsistent properties (e.g. being 5 inches long and being 7 inches long), then (1) Fa, (2) Gb, and (3) a=b cannot all be true (something cannot be both 5 and 7 inches long). How might one respond?

On the one hand, one might hold fast to the principle that no genuine unity can have inconsistent properties, and conclude that no genuine unity in fact persists through change at all. No numerical identity without qualitative identity. Thus, we make the distinction between enduring unities and perduring unities, and insist that objects persist in virtue of perduring (through a succession of momentary genuine unities appropriately related to one another), not in virtue of enduring.

On the other hand, we might take seriously the idea that time is doing some important work here, and allow that while a genuine unity cannot have inconsistent properties at any one time, having inconsistent properties at different times might be tolerated somehow (in a way that is to be explicitly specified). So, we allow for the possibility of numerical identity in the absence of qualitative identity. Since numercal unity cannot be grounded in qualitative identity on this route, we must ground it in something else, and there are two prominent options. One might restrict the class of properties that are required to remain the same in order for the numerical identity of the thing to be preserved: the *essential* properties do not change, no object has associated with it a set of inconsistent essential properties, not even over time. As for the accidental properties, we require that these are consistent at any one time, but we do not to care whether they contain inconsistencies over time. Alternatively, one might claim that numerical identity over time is *independent* of sameness of properties over time: we appeal to haecceities to ground numerical identity over time, and we don't care about any inconsistencies in properties over time (although we continue to require that an object's properties be consistent at any one time). This allows for genuine unities which persist in virtue of *enduring*.

There are good reasons for philosophers of physics to be sceptical about both essentialism and haecceitism, which appears to leave "no numerical identity without qualitative identity" as a feature of our account of unity, and consequently perdurantism as our account of change, as the only option. But the law-constitutive approach reveals an alternative. The law-constitutive approach offers a principle of unity in virtue of which a thing remains the very same thing over time and through change of properties. It does so *not* by appeal to haecceities, *nor* by appeal to essential properties, but by specifying the relations that must hold between the states of the thing at different times. You might think: okay, but this view is compatible with both perdurantism and endurantism: for the perdurantist the laws specify the relationship between the successive momentary genuine unities for the endurantist the laws specify the relationship between successive states of a single genuine unity. But this isn't right. Here's the crucial

question: what are the perdurantist's "momentary genuine unities" that are supposedly tied together by the laws? *In virtue of what* are these – the things that are tied together – themselves *genuine unities*? If the genuine unity is grounded in qualitative identity, give me an argument why I should accept this view. If it's grounded in something else, tell me what. In the absence of an answer to this, I think that the law-constitutive approach gives us an argument in favor of endurantism as against perdurantism, because according to the law-constitutive person the very principle that grounds the unity of a thing has *as one of its consequences* rules by which such a unity can undergo qualitative change.

This is a key point, and for that reason I want to emphasize it a little further. In generating the *prima facie* puzzle about change, we had to write down "a=b". But in order to write this down, we have to presuppose that our things labelled by "a" and "b" are genuine unities, and we need an account of what grounds that unity. We can't take unity as brute, at least not without saying why the worries of the seventeenth century philosophers were misplaced.[1] So, in the absence of a principle of unity, suitably argued for, the perdurantist is at a disadvantage as compared to the endurantist. The law-constitutive approach offers a principle of unity which provides numerical identity without qualitative identity, and it provides an account of what it is for a genuine unity to undergo change. It's an approach that arose within attempts to construct a physics and a metaphysics of things by two giants of this enterprise: Descartes and Newton. What it gives us is a reason to prefer endurantism, and from here it is perhaps a short step to presentism, for while both endurantism and perdurantism are compatible with both presentism and four-dimensionalism, most metaphysicians think there is a more natural fit between

[1] In our theorizing about the world, *objects* should not be taken as primitives. As Saunders (2003) argues, we have access (at least in physics) first of all not to objects, but to their properties and relations, and (for the purposes of physics at least) identity of objects needs to be defined in these terms, not taken as primitive. For example, Della Rocca (forthcoming) diagnoses an apparent stand-off between endurantism and perdurantism, and then argues against endurantism on the grounds that the endurantist must take persistence as primitive. The implicit assumption Della Rocca makes is that objects are to be taken as primitives: in the law-constitutive approach, objects are not primitive, and neither is persistence. Thus, endurantism escapes Della Rocca's argument.

endurantism and presentism and between perdurantism and four-dimensionalism. If that's right, then the law-constitutive approach to unity and change favors presentism.

This leaves us with a tension. On the one hand, considerations arising out of space-time theory push strongly towards the block universe. On the other hand, consideration arising out of Newton's physics lead to a form of presentism. In the final sections of my paper I attempt to remove this tension, and in so doing I further elaborate the alternative form of presentism that I take to be both (a) an empirical hypothesis, and (b) compatible with special relativity.

6. Space and time as an epistemic principle of unity

At the end of section 3, I claimed that we don't *have* to use spacetime to play the role of an ontological principle of unity. In section 4 I offered a sketch of an alternative, and in section 5 I showed how this alternative favors a version of presentism. If we are to take this route (and I fully concede that it stands in great need of significant further elaboration), then we should re-visit the status of spacetime. If it no longer serves as a principle of ontological unity, what role does it play? Why do we set out a big arena of space and time when we're doing physics? The answer, I think, is that spacetime plays the role of an *epistemic* principle of unity, as follows. In mechanics, we want to know what the outcome of a collision will be, before it happens, based on knowledge of events prior to the collision. Quite generally, one thing we're doing is trying to extend our knowledge of events to times and places distant from the here and now. Space&time play a theoretical role, as we try to extend our epistemic reach beyond the here-now, stitching our predictions together into a single whole. Thinking of things in this way, space&time provide an *epistemic* principle of unity: they provide the framework in which we organize our knowledge of the not-here and/or not-now.

There is no necessary inference from this epistemic role for spacetime to the view that spacetime is an *ontological* principle of unity. Thus, if we take this route, we should revise our understanding of (P2). In being committed to special relativity as a complete account of spatiotemporal structure, we do not thereby automatically ontologize this structure: we recognize its epistemic status and we do not make any direct inference from that to any ontological commitments.

Instead, we adopt the approach sketched in section 4, according to which the laws provide the principle of ontological unity. Thus, whatever spatiotemporal ontological commitments we have must come from paying attention to the details of the dynamical laws of matter. Matter is spatiotemporal, but it's not *in* space and time in the sense that space and time provide an ontological principle of unity for what there is.[2]

According to this approach, the dynamical laws ground the unity of a thing, and the spatiotemporal extent of that thing is whatever size it needs to be in order to sustain the dynamically characterized thing in question. Once this approach is adopted, (P2) means two things:

(1) We take special relativity seriously as an epistemic principle of unity: this is the best way of organizing our knowledge that reaches beyond the here-now.

(2) Ontologically, what there is is grounded in the dynamical laws, and these include the spatiotemporal characteristics of things. If things are spatiotemporally extended, then special relativity tells us that within that spacetime region there are no purely spatial or purely temporal relations. Things "occupy" time just as they "occupy" space: by existing as a unity that is spatiotemporally extended. Notice that the spatiotemporal extent of the sufficient dynamical ground of a given unity might turn out to be much, much smaller than the abstract spatiotemporal structure within which the evolution of that unity is

[2] Ontologically, space and time are not independent of bodies. But this is not relationism either, so we move immediately to a middle way between ab space and time versus relationism. (This is a second reason for the superiority of the dynamical approach.)

completely described, or it might be the size of the block universe.[3] This is an empirical matter, something to be settled by the progress of science. I will return to this point in a moment.

7. An alternative "presentism"

What becomes of presentism on this view? (P1) will need to be re-written. The reason is that, on the approach being developed here, we are not going to start from Minkowkski spacetime when we do ontology. Hence, the presentist should not use "now" as the grounds of what is real: she should use dynamics instead. *The present is a spatiotemporal region of whatever size is necessary to sustain the dynamical system in question*.[4] If we take this route, then we will not be driven by our considerations of the structure of Minkowski spacetime to the conclusion that the present is merely a point in Minkowski spacetime, and that therefore this is all that is real. In other words, since we are not using a "now" derived from the structure of Minkowksi spacetime to ground what is real, we don't end up at pluralistic extreme solipsism.

It might turn out that, for whatever system we try to consider, the size of the spatiotemporal region necessary to sustain it is the entire history of the universe, encompassing all that ever has been and ever will be. In other words, the only dynamical system that there is, is this entirety, and there are no genuine subsystems of the universe. If that turns out to be the case, then this version of presentism is defeated and the block universe triumphs. But notice that this is an empirical matter, something to be decided by consideration of the details of physics, and perhaps science more generally. Quantum entanglement might give us good reason to think that this is true, but if these consideration lead to the

[3] I can be made of things whose dynamically sustained regions of spacetime and much smaller than mine, so they come into and go out of existence on much shorter timescales than I do. That seems fine to me. (There's always something existing whenever I exist – it's hard to say this right, I can see this is going to be an interesting challenge to express this thesis!)

[4] There are no determinate *ontological* spatial or temporal relations within that region, and that system stands in no determinate *ontological* spatial or temporal relations to any other system. All the ontology is carried by the dynamics, and we frame the dynamics spatiotemporally, but ontologically the dynamics requires no such spatiotemporal underpinning.

defeat of presentism then they do so via a different route than Einstein's treatment of simultaneity. Notice also, however, that this issue is not yet settled.

So let's begin from the position that there are genuine subsystems. This means that the size of the spatiotemporal region required to sustain the system is less than the entire block universe, and so the present is (at least in the first instance) local, not global. This "local now" does not lead to solipsism, however, because it is not the ground of what is real. The grounds of what is real is the dynamics, and we belong to the same world as whatever we interact with, and the rest of that world is as real as we are. The dynamics grounds the unity of what there is, both of the parts and of the whole (consisiting of interacting parts), and this is what prevents the presentist from becoming a solipsist. Does my son exist relative to me, when he is in London and I am here? Of course. There are plenty of interactions going on that link us. Is there a determinate fact of the matter about what he is doing *right now*? No.

Clearly, there is a lot of work to be done in filling out exactly what this position says. But one thing we can do straight away is reformulate presentism such that it doesn't ground the reality of what exists in spacetime. Here's an attempt at a better (P1):

(P1*) For each and every thing, that thing exists only presently, where the spatiotemporal extent of that "present" is dependent on dynamics, and it is something to be determined empirically

This is a version of presentism that endorses (P2).

8. Conclusion

That we can systematize things in a global spatiotemporal framework is surely an interesting fact, and you might want to ontologize the overall framework above and beyond the dynamics. You can if you

want to. My point has been that you don't have to, and that if you're going to you need to say why adopting spacetime as the ground of ontological unity is better than using the dynamical laws. Should you choose to adopt dynamical laws as the ground of unity, an alternative version of presentism emerges, and one which (unlike the version grounded in spacetime structure) has yet to be empirically falsified. A great deal rests on whether there are genuine subsystems of the universe, and that's something that we can find out only through empirical enquiry.

References

Della Rocca, Michael. Forthcoming.

DiSalle, Robert. 2006. *Understanding Space-Time: The philosophical development of physics from Newton to Einstein*. Cambridge: Cambridge University Press.

Saunders, Simon. 2003. Physics and Leibniz's Principles. *Symmetries in Physics: Philosophical Reflections*, K. Brading and E. Castellani, eds., Cambridge University Press.

Stein, Howard. 1968. On Einstein-Minkowski Space-Time. *Journal of Philosophy* 65(1):5-23.

4,380 words

Values in Science beyond Underdetermination and Inductive Risk

Matthew J. Brown
Center for Values in Medicine, Science, and Technology
The University of Texas at Dallas
mattbrown@utdallas.edu

July 12, 2012

Abstract

The thesis that the practice and evaluation of science requires social value-judgment, that good science is not value-free or value-neutral but value-laden, has been gaining acceptance among philosophers of science. The main proponents of the value-ladenness of science rely on either arguments from the underdetermination of theory by evidence or arguments from inductive risk. Both arguments share the premise that we should only consider values once the evidence runs out, or where it leaves uncertainty; they adopt a criterion of *lexical priority of evidence over values*. The motivation behind lexical priority is to avoid reaching conclusions on the basis of wishful thinking rather than good evidence. *The problem of wishful thinking* is indeed real—it would be an egregious error to adopt beliefs about the world *because* they comport with how one would prefer the world to be. I will argue, however, that giving lexical priority to evidential considerations over values is a mistake, and unnecessary for adequately avoiding the problem of wishful thinking. Values have a deeper role to play in science than proponents of the underdetermination and inductive risk arguments have suggested.

1 Introduction

This paper is part of the larger project of trying to understand the structure of values in science, i.e., the role of values in the logic of scientific practice. This is

distinct from the project of strategic arguments that try to establish *that* science is value-laden while assuming premises of the defenders of the value-free ideal of science. It is becoming increasingly hard to deny that values play a role in scientific practice—specifically non-epistemic, non-cognitive, or contextual values, e.g., moral, political, and aesthetic values (I will use the term "social values" to refer to such values in general). What is less clear is what parts of scientific practice require values or value-judgments. This is not primarily a historical or sociological question, though historical and sociological data is frequently brought to bear. Ultimately it is a *normative* question about the role that value-judgments *ought* to play in science; it is a question about the proper *ideal* of scientific practice. As such, we must consider both ethical questions about how the *responsible* conduct of science requires value-judgment and epistemological questions about how the *objectivity* and *reliability* of science is to be preserved.

There are a number of phases of inquiry where values might play a role: (1) in determing the value of science itself and (2) the research agenda to be pursued, (3) in framing the problem under investigation and (4) the methods of data collection and characterization, (5) in choosing the hypothesis, explanation, or solution to propose, (6) in the testing or certification of a proposed solution, and (7) in choices about application and dissemination of results. Various accounts have allowed values in some stages while excluding it in others, or have argued for specific limits on the role for values at each stage. In this paper, I will focus on the testing phase, where theories are compared with evidence and certified (or not) as knowledge, as this is the most central arena for discussion value-free vs. value-laden science. Traditionally, philosophers of science have accepted a role for values in practice because it could be marginalized into the "context of discovery," while the "context of justification" could be treated as epistemically pure. Once we turn from the logical context of justification to the actual context of certification[1] in practice, the testing of hypotheses within concrete inquiries conducted by particular scientists, we can no longer ignore the role of value-judgments.

There are two main arguments in the literature for this claim: *the error argument* from inductive risk and *the gap argument* from the underdetermination of theory by evidence. While both of these arguments have been historically very important and have successfully established important roles for values in science, they share a flawed

[1] I use "context of certification" following Kitcher (2011), as referring to actual practices of acceptance. While I won't emphasize it in this paper, I also follow Kitcher in thinking that certification is a *social* practice that results in accepting a result as part of *public* knowledge (as opposed to merely individual belief).

premise, the *lexical priority of evidence over values*.[2] While this premise serves an important aim, that of avoiding the *problem of wishful thinking*, I will argue that there are several problems with this premise. We should seek an alternative ideal for science that provides a role for values at a more fundamental level and broader scope, but nevertheless preserves an important feature of science: the ability to surprise us with new information beyond or contrary to what we already hope or believe to be true.

2 Underdetermination: The Gap Argument

Underdetermination arguments for the value-ladenness of science extend Duhem's and Quine's thoughts about testing and certification. The starting point for this argument may be the so-called Duhem-Quine Thesis (or Duhem-Neurath-Quine Thesis (Rutte, 1991, p. 87)) that no hypothesis can be tested in isolation because of the need for auxiliary assumptions in order for theories to generate testable hypotheses. This is generally taken to imply that no theory can be definitively falsified by evidence, as the choice between rejecting the theory, altering the background assumptions, or even (though more controversially) rejecting the new evidence itself as faulty is *underdetermined* by each new item of evidence—call this "holist underdetermination" (Stanford, 2009).

Another form of underdetermination—"contrastive underdetermination" (*ibid.*)—depends on the choice between identically confirmed rival hypotheses. As all of the evidence available equally supports either hypothesis in such cases, that choice is underdetermined by the evidence. If the evidence we're talking about is just all the evidence we have available to us at present, then we have *transient* underdetermination, which might be relatively temporary or might be a *recurrent* problem. If instead the choice is underdetermined by *all possible* evidence, we have *permanent* underdetermination and the competing theories or hypotheses are *empirically equivalent*. The global underdetermination thesis holds that permanent underdetermination is ubiquitous in science, applying to all theories and hypotheses.[3]

The many forms of underdetermination argument have in common the idea that some form of *gap* exists between theory and observation. Feminists, pragmatists,

[2]Strictly speaking, both arguments can be taken as *strategic* arguments, compatible with any positive approach to the role of values in scientific inquiry. For the purposes of this paper, I will instead take the arguments as attempts to articulate a positive ideal. The gap and error arguments are perfectly serviceable as strategic arguments.

[3]For discussion of forms of underdetermination, see Kitcher (2001); Magnus (2003); Stanford (2009); Intemann (2005); Biddle (2011).

and others have sought to fill that gap with social values, or to argue that doing so does not violate rational prescriptions on scientific inference. Call this *the gap argument* for value-laden science (Intemann, 2005; Elliott, 2011). Kitcher (2001) has argued that *permanent* or *global* underdetermination is needed to defeat the value-free ideal of science, and these forms of underdetermination are much more controversial. Transient underdetermination, on the other hand, is "familiar and unthreatening," even "mundane"(Kitcher, 2001, p. 30-1)

Kitcher is wrong on this point; *transient underdetermination* is sufficient to establish the value-ladenness of scientific practice (Biddle, 2011). What matters are decisions made in practice by actual scientists, and at least in many areas of cutting edge and policy-relevant science, transient underdetermination is pervasive. Perhaps it is the case that in the long run of science (in an imagined Peircean "end of inquiry") all value-judgments would wash out. But as the cliché goes, in the long run we're all dead; for the purposes of this discussion, what we're concerned with is decisions made *now*, in the actual course of scientific practices, where the decision to accept or reject a hypothesis has pressing consequences. In such cases, we cannot wait for the end of inquiry for scientists to accept or reject a hypothesis, we cannot depend on anyone else to do it, and we must contend with uncertainty and underdetermination. Actual scientific practice supports this—scientists find themselves in the business of accepting and rejecting hypotheses in such conditions.

So what is the role for social values under conditions of transient underdetermination? Once the existing evidence is in, a gap remains in definitively determining how it bears on the hypothesis (holist case) or which competing hypothesis to accept (contrastive case). In this case, it can be legitimate to fill the gap with social values. For example, among the competing hypotheses still compatible with all the evidence, one might accept the one whose acceptance is likely to do the most good or the least harm. E.g., in social science work involving gender or race, this might be the hypothesis compatible with egalitarianism.

A common response is that despite the existence of the gap, we should ensure that no social values enter into decisions about how to make the underdetermined choice (e.g., whether or not to accept a hypothesis). Instead, we might fill the gap with more complex inferential criteria (Norton, 2008) or with so-called "epistemic" or "cognitive" values (Kuhn, 1977; Laudan, 1984). Proponents of the gap argument have argued that this at best pushes the question back one level, as choices of epistemic criteria or cognitive values (Longino, 2002, p. 185), and application of cognitive values itself may not be entirely determinate (Kuhn, 1977). Ensuring that no values actually enter into decisions to accept or reject hypotheses under conditions of transient underdetermination may turn out to be *impossible* (Biddle, 2011).

Another attempt to avoid a role for social value-judgments—withholding judgment until transient underdetermination can be overcome or resolved by application of cognitive factors along—is *unreasonable* or *irresponsible* in many cases, e.g. where urgent action requires commitment to one or another option (ibid.).[4]

What distinguishes *legitimate* from *illegitimate* uses of values to fill the gap is a matter of controversy, sometimes left unspecified. With some exceptions,[5] underdeterminationists insist that values only come into play in filling the gap (e.g., Longino, 1990, p. 52, 2002, p. 127; Kourany, 2003).

3 Inductive Risk: The Error Argument

While underdeterminationist arguments for values in science are probably more well known, and may have a history going back a paper of Neurath's from 1913 (Howard, 2006), the *inductive risk argument* for values in science is older still, going back to William James' (1896) article "The Will to Believe."[6] Heather Douglas has revived Rudner's (1953) and Hempel's (1965) version of the argument for the value-ladenness of science. In simplified form, the argument goes like this:

In accepting or rejecting hypotheses, scientists can never have complete certainty that they are making the right choice—uncertainty is endemic to ampliative inference. So, inquirers must decide whether there is *enough* evidence to accept or reject the hypothesis. What counts as *enough* should be determined by how *important* the question is, i.e., the *seriousness* of making a mistake. That importance or seriousness is generally (in part) an *ethical* question, dependent on the ethical evaluation of the consequences of error. Call this argument for the use of value-judgments in science from the existence of inductive risk *the error argument* (Elliott, 2011).

According to the error argument, the main role for values in certification of scientific hypotheses has to do with how much uncertainty to accept, or how strict to make your standards for acceptance. In statistical contexts, we can think of this as the trade-off between *type I* and *type II* error. Once we have a fixed sample size (and assuming we have no control over the effect size), the only way we can decrease the probability that we wrongly reject the null hypothesis is to increase the probability

[4] Proponents of the inductive risk argument make a similar point.

[5] These exceptions either use a somewhat different sort of appeal to underdetermination than the gap argument, or they use the gap argument as a strategic argument. One example is the extension of the Quinean web of belief to include value-judgments (Nelson, 1990), discussed in more detail below.

[6] This connection is due to P.D. Magnus (2012), who refers to the inductive risk argument as the "James-Rudner-Douglas or JRD thesis" for reasons that will become immediately apparent.

that we wrongly accept the null hypothesis (or, perhaps more carefully, that we fail to reject the null hypothesis when it is in fact false), and vice versa. Suppose we are looking for a causal link between a certain chemical compound and liver cancers in rats,[7] and you take H_0 to be no link whatsoever. If you want to be absolutely sure that you don't say that the chemical is safe when it in fact is not (because you value safety, precaution, welfare of potential third parties), you should decrease your rate of type II errors, and thus increase your statistical significance factor and your rate of type I errors. If you want to avoid "crying wolf" and asserting a link where none exists (because you value economic benefits that come with avoiding overregulation), you should do the reverse.

Douglas emphasizes at length that values (neither social nor cognitive values) should not be taken as *reasons* for accepting or rejecting a hypothesis, reasons on a par with or having the same sort of role as *evidence* in testing.[8] This is an impermissible *direct* role for values. In their permissible *indirect* role, values help determine the rules of scientific *method*, e.g., decisions about how many false positives or false negatives to accept. Values are not reasons guiding belief or acceptance; they instead guide decisions about how to manage uncertainty.[9]

Rudner (1953) anticipated the objection that scientists should not be in the business of accepting or rejecting hypothesis, but rather just indicating their probability (and thus not having to make the decision described above). This response wrongly assumes that inductive risk only occurs at the final step of certification; in reality, this gambit only pushes the inductive risk back a step to the determination of probabilities. Furthermore, the pragmatic signal that accompanies a refusal to assent or deny a claim in practical or policy circumstances may be that the claim is far more questionable that the probabilities support. Simply ignoring the consequences of error—by refusing to accept or reject, by relying only on cognitive values, or by choosing purely conventional levels for error—may be *irresponsible*, as scientists like anyone else have the moral responsibility to consider the foreseeable consequences of their action.

[7] Douglas (2000) considers the actual research on this link with dioxin.

[8] Strictly speaking, this is an extension of the error argument, and not all who accept the argument (especially for strategic purposes) need accept this addition.

[9] In Toulmin's (1958) terms, values cannot work as grounds for claims, but they can work as backing for warrants.

4 A Shared Premise

These two arguments against the value-free ideal of science share a common premise. The gap argument holds that values can play a role in the space fixed by the evidence; if the gap narrows (as it would with transient underdetermination), there are fewer ways in which values can play a role, and *if* the gap could ever be close, the conclusion would be value-free. (An exception are those views that add values into the radically holistic interpretation of Quine's web of belief, such that values, theories, and evidence are all equally revisable in the light of new evidence.) The inductive risk argument allows values to play a role in decisions about how to manage uncertainty—not directly by telling us which option to pick, but indirectly in determining how much uncertainty is acceptable.

Both arguments begin from a situation where the evidence is fixed and take values to play a role in the space that is left over. The reason that values must play a role is that uncertainty remains once the evidence is in. In a relatively weak version of this argument, social values fill in the space between evidence and theory because something has to, so it might as well be (and often is) social values. In more sophisticated versions, we must use social values to fill the gap because of our general moral obligation to consider the foreseeable consequences of our actions, including the action of accepting a hypothesis. The arguments of these two general forms all assume the *lexical priority of evidence over values*. The premise of lexical priority guarantees that even in value-laden science, values do not compete with evidence when the two conflict. This is often defended as an important guarantor of the objectivity or reliability of the science in question.

5 Why Priority?

Why do proponents of value-laden science tend to be attracted to such a strict priority of evidence over values? Perhaps some such restriction is required in order to guarantee the *objectivity* of science. In order for our science to be as objective as possible, maybe it has to be as value-free as possible (though this may not be very value-free at all). That is, we want as much as possible to base our science on the evidence because evidence lends objectivity and values detract from it. Even if this view of objectivity were right, however, it would be a problematic justification for opponents of the value-free ideal of science to adopt. With arguments like the gap and inductive risk arguments, they mean to argue that values and objectivity are not in conflict *as such*. It would thus create a serious tension in their view if one premise depended on such a conflict. If it is really *objectivity* that is at stake in adopting

lexical priority, we need a more nuanced approach.

I think the central concern concern is that value judgments might "drive inquiry to a predetermined conclusion"(Anderson, 2004, p. 11), that inquirers might rig the game in favor of their preferred values. As Douglas (2009) puts it, "Values are not evidence; wishing does not make it so"(p. 87). In other words, a core value of science is its ability to *surprise* us, to force us to revise our thinking. Call the threat of values interfering with this process *the problem of wishful thinking*.

Lexical priority avoids this problem insofar as what we value (which involves the way we desire the world to be) is only a consideration *after* we take all of the evidence (which fixes the way the world is) into account. In Douglas's more nuanced approach, even once the evidence is in, social values (and even most cognitive values) are not allowed to be taken *directly* as reasons to believe anything; they only act as reasons for accepting a certain amount of evidence as "enough."

An alternative explanation may be that the adoption of lexical priority has *rhetorical value*.[10] Suppose, along with the defenders of the value-free ideal, that there is such a thing as *objective evidence* which constrains belief. Even so, there is (at least transient) underdetermination, and a gap that must bridged by social values. Thus not only is the value-free ideal impossible to realize, it may lead to unreasonable and irresponsible avoidance of the role for values in filling the gap. Such an argument can undermine the value-free ideal and establish that there is a major role for values in science, and in the context of these goals, I freely admit that this can be a worthwhile strategy. But as we turn instead to the positive project of determining more precisely the role(s) of values in the logic of scientific practice, the premises of such an immanent critique are unfit ground for further development. We no longer need to take the premises of our opponents on board, and we may find that they lead us astray.

While following the basic contours of my argument so far, one might object to characterizing of evidence as "prior" to values.[11] What the gap and inductive risk arguments purport to show is that there is always some uncertainty in scientific inference (perhaps, for even more basic reasons, in all ampliative inference), and so there will always be value-judgments to be made about when we have enough evidence, or which among equally supported hypotheses we wish to accept, etc. The pervasive need for such judgments means that value-freedom does not even make sense as a limiting case; both values and evidence play a role, and neither is prior to the other. This mistakes the sense of "priority" at work, however. Where priority matters is what happens when values and evidence conflict; in such circumstances,

[10] *Note redacted for purposes of anonymous review.*

[11] *Note redacted for purposes of anonymous review.*

lexical priority means that evidence will always trump values. In Douglas's stronger version of lexical priority, values allow you to determined what level of evidence you need to accept a hypothesis ($p = 0.05$ or $p = 0.01$ or...), but they cannot give you a *reason* to reject the hypothesis,[12] no matter what.

6 Problems with Priority

The versions of the gap and inductive risk arguments that presuppose the lexical priority of evidence make two related mistakes. First, they require a relatively uncritical stance towards the status of evidence *within* the context of certification.[13] The lexical priority principle assumes that in testing, we ask: given the evidence, what should we make of our hypothesis? Frame this way, values only play a role at the margins of the process.

This is a mistake, since evidence can turn out to be bad in all sorts of ways: unreliable, unrepresentative, noisy, laden with unsuitable concepts and interpretations, or irrelevant for the question at hand; the experimental apparatus could even have a cord loose. More importantly, we may be totally unaware of why the evidence is bad; after all, it took a great deal of ingenuity on the part of Galileo to show why the tower experiment didn't refute Copernicus, and it took much longer to deal with the problem of the "missing" stellar parallax. While some epistemologists stick to an abstract conception of evidence according to which evidence is itself unquestionable, reflection on cases like this has lead many philosophers of science to recognize that we can be skeptical about particular pieces or sets of evidence based on its clash with hypotheses, theories, or background assumptions that we have *other* good reasons to hold on to. As critics of strict falsificationism and empiricism have shown, we already have reason to adopt a more egalitarian account of the process of testing and certification, *independent* of the question about the role of values. We might get off to a better start if we thought about how to fit values into this sort of picture of testing.

[12] It seems possible that we could use our extreme aversion to some hypothesis to raise the required level of certainty so high as to be at least *practically* unsatisfiable by human inquirers, and so in effect rule out the hypothesis on the basis of values alone while remaining in the indirect role. While it isn't clear how to do it, it seems to be that Douglas means to rule this sort of case out as well.

[13] As Douglas (2009) makes clear, she does not take the status of evidence as unproblematic *as such*. But any issues with the evidence are to be taken into account by prior consideration of values in selection of methods and characterization of data. It would seem that value judgments in the context of certification cannot be a reason to challenge the evidence itself. The following points are intended to show that this restriction is unreasonable.

Second, the attitude about values that lexical priority takes reduces the idea of value judgment to merely expression of *preferences* rather than *judgment* properly so called—in effect, they deny that we can have good reasons for our value judgments. It is crucial to distinguish between values or valuing and value judgments or evaluations (Dewey, 1915, 1939; Welchman, 2002; Anderson, 2010). Valuing may be the mere expression of a preference, but value judgments are reflective decisions about values, and properly speaking must be made on the basis of reasons (and judgments can be better or worse because they are made on the basis of good and bad reasons). Value judgments may even be open to a certain sort of empirical test, because they hypothesize relationships between a state or course of action to prefer and pursue and the desirability or value of the consequences of pursuing and attaining them (Dewey, 1915; Anderson, 2010). Value judgments say something like "try it, you'll like it"—a testable hypothesis (Anderson, 2010). The evidence by which we test value judgments may include the emotional experiences that follow on adopting those values (Anderson, 2004).

If value judgments are judgments properly so called, adopted for good reasons, subject to certain sorts of tests, then it is unreasonable to treat them in the manner required by the lexical priority of evidence. Just as the good (partly empirical) reasons for adopting a theory, hypothesis, or background assumption can give us good reasons to reinterpret, reject, or maybe even ignore evidence apparently in conflict with them (under certain conditions), so too with a good value judgment. If evidence and values pull in opposite directions on the acceptance of a hypothesis, then we should not always be forced to follow the (putative) evidence.

7 Avoiding Wishful Thinking without Priority

If we reject the lexical priority assumption and adopt a more egalitarian model of testing, we need to adopt an alternative approach that can avoid the problem of wishful thinking.

(An alternative principle to lexical priority is *the joint necessity of evidence and values*, which requires joint satisfaction of epistemic criteria and social values. This is the approach taken by Kourany (2010). On such a view, neither evidence nor values takes priority, but this principle leaves open the question of what to do when evidence and values clash. One option is to remain *dogmatic* about both epistemic criteria and social values, and to regard any solution which flouts either as a failure, which appears to be Kourany's response.

Alternatively, we can adopt **the rational revisability of evidence and values** in addition to joint necessity and revisit and refine our evidence or values. On this

principle, both the production of evidence and value formation are recognized as rational but fallible processes, open to revision. Such a view might include the radical version of Quinean holism which inserts values into the web of belief. The adoption of these two principles alone does not prevent wishful thinking, but adding some basic principles like *minimal mutilation* may overcome the problem. (cf. Kitcher, 2011)

Instead of Quinean holism, we might instead adopt a form of **pragmatist functionalism about inquiry** (Brown, 2012) which differentiates the functional roles of evidence, theory, and values in inquiry. This retains the idea that all three have to be coordinated and that each is revisable in the face of new experience, while introducing further structure into their interactions and According to such an account, not only must evidence, theory, and values fit together fit together in their functional roles, they must do so in a way that *actually* resolves the problem that spurred the inquiry.

8 Conclusion

The lexical priority of evidence over values is an undesirable commitment, and unnecessary for solving the problem it was intended to solve. The key to the problem of wishful thinking is that we not predetermine the conclusion of inquiry, that we leave ourself open to surprise. The real problem is not the insertion of values, but *dogmatism* about values (Anderson 2004). Rather than being the best way to avoid dogmatism, the lexical priority of evidence over values coheres best with a *dogmatic* picture of value judgments, and so encourages the illegitimate use of values. A better account is one where values and evidence are treated as mutually necessary, functionally differentiated, and rationally revisable components of certification. Such an account would allow that evidence *may* be rejected because of lack of fit with a favored hypothesis and compelling value-judgments, but *only* so long as one is still able to effectively solve the problem of inquiry.

References

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia 19*(1), 1–24.

Anderson, E. (2010). Dewey's moral philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2010 ed.).

Biddle, J. (2011, October). Transient underdetermination, value freedom, and the epistemic purity of science. Unpublished manuscript.

Brown, M. J. (2012, Fall). John Dewey's Logic of Science. *HOPOS: The Journal of the International Society for the History of Philosophy of Science 2*(2).

Dewey, J. (1915). The logic of judgments of practice. In J. A. Boydston (Ed.), *The Middle Works, 1899–1924*, Volume 8. Carbondale: Southern Illinois University Press.

Dewey, J. (1939). *Theory of Valuation*. In J. A. Boydston (Ed.), *The Later Works, 1925–1953*, Volume 13. Carbondale: Southern Illinois University Press.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science 67*(4), 559–579.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, K. C. (2011). *Is a little pollution good for you?: incorporating societal values in environmental research*. Environmental ethics and science policy series. New York: Oxford University Press.

Hempel, C. G. (1965). Science and human values. In *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, pp. 81–96. New York: The Free Press.

Howard, D. (2006). Lost wanderers in the forest of knowledge: Some thoughts on the discovery-justification distinction. In J. Schickore and F. Steinle (Eds.), *Revisiting Discovery and Justification: Historical and Philosophical Perspectives on the Context Distinction*, pp. 3–22. Dordrecht: Springer.

Intemann, K. (2005). Feminism, underdetermination, and values in science. *Philosophy of science 72*(5), 1001–1012.

James, W. (1896). The will to believe. *The New World 5*, 327–347.

Kitcher, P. (2001). *Science, Truth, and Democracy*. Oxford University Press.

Kitcher, P. (2011). *Science in a democratic society*. Amherst, N.Y.: Prometheus Books.

Kourany, J. A. (2003). A philosophy of science for the twenty-first century. *Philosophy of science 70*(1), 1–14.

Kourany, J. A. (2010). *Philosophy of science after feminism*. Oxford Univ Pr.

Kuhn, T. S. (1977). *Objectivity, Value Judgment, and Theory Choice*, pp. 320–39. Chicago: University of Chicago Press.

Laudan, L. (1984). *Science and values: the aims of science and their role in scientific debate*. Berkeley: University of California Press.

Longino, H. E. (1990). *Science as social knowledge: values and objectivity in scientific inquiry*. Princeton, N.J.: Princeton University Press.

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Magnus, P. (2003). *Underdetermination and the Claims of Science*. Ph. D. thesis, University of California, San Diego.

Magnus, P. (2012). What scientists know is not a function of what scientists know. In *PSA 2012*.

Nelson, L. H. (1990). *Who knows: from Quine to a feminist empiricism*. Philadelphia: Temple University Press.

Norton, J. (2008). Must evidence underdetermine theory. *The challenge of the social and the pressure of practice: Science and values revisited*, 17–44.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science 20*(1), 1–6.

Rutte, H. (1991). The Philosopher Otto Neurath. In T. E. Uebel (Ed.), *Rediscovering the Forgotten Vienna Circle: Austrian Studies on Otto Neurath and the Vienna Circle, Kluwer Academic Publishers, Dordrecht*, pp. 81–94. Kluwer Academic Publishers.

Stanford, K. (2009). Underdetermination of scientific theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2009 ed.).

Toulmin, S. (1958). *The Uses of Argument*. Cambridge, U.K.: Cambridge University Press.

Welchman, J. (2002). Logic and judgments of practice. In F. T. Burke, D. M. Hester, and R. B. Talisse (Eds.), *Dewey's logical theory: new studies and interpretations*. Vanderbilt Univ Press.

POPPER'S CONTRIBUTION TO THE PHILOSOPHICAL STUDY OF ARTIFACTS

1. Introduction

Research on the nature and function of artifacts has provided one of the richest areas of discussion in contemporary philosophy of technology as can be witnessed most notably in, for instance, the works of Don Ihde (1990), Peter-Paul Verbeek (2005), Peter Kroes and Anthony Meijers (2006). Some promising developments regarding the philosophical study of artifacts can be noted outside of this field as well. Signs of an emerging interest in artifacts are visible in fields like philosophy of science (Hacking 1983; Ihde 1991) and contemporary metaphysics. Quite interestingly, one of the most striking features about discussions of artifacts in recent metaphysics (van Inwagen 1990; Wiggins 2001; Elder 2004) is the tacit denial of their existence on various grounds. Important exceptions to this contemporary 'eliminativist' trend include Randall Dipert's (1993) study of artifacts, the works of Baker (2004, 2007) and Thomasson (2003, 2009).

This increasing numbers of research publications notwithstanding, a detailed systematic, critical study of artifacts is still in its infancy. Philosophers generally prefer to occupy themselves with words and ideas and do not tend to devote serious attention to various tools, appliances, or other technological devices. Quite predictably, at the level of theory and reflection these technological products have been relatively neglected by the philosophers.[1] However, if the task of philosophy, as is commonly understood, is to help us make sense of the human world and to deal with the most fundamental traits of reality (which includes not only the objects of natural sciences but indeed of those studied by the social, human and technological sciences as well), then the question of how to secure the significance of artifacts in philosophical investigation is more pressing than ever.

The principal motivation behind this inquiry is to display the versatility of Popper's thesis of three worlds (1972, 1977, 1979, and 1982) in the analysis of issues related to the ontological status and character of artifacts. Strange to say, despite being discussed over years and hit with numerous criticisms (Carr 1977; Currie 1978; Cohen 1980), it is still little known that Popper's thesis has an important bearing on the philosophical characterization of technical

1This, however, does not imply that material things have been totally neglected by the human and social sciences. An interest in things has its own long tradition, including the history of art, archaeology and what is often referred to as 'classical philosophy of technology'.

artifacts[2]. In consequence, his key perspectives on the reality, autonomy, and ontological status of artifacts are rarely taken into consideration by scholars known to be engaged in the study of artifacts.[3] In this paper I intend to address this unfortunate oversight.

This paper consists of two main sections. The first section attempts to present a critical exposition of Popper's account of reality and (partial) autonomy of artifacts. Recent discussions about the longstanding distinction between natural objects[4] and artifacts are brought up and the relevance of Popper's pluralistic thesis to this debate is pointed out. In addition, attention is drawn towards how to read his notion of the autonomy of artifacts. The primary emphasis of the second section is the ontological position of artifacts. Two separate arguments are posed to challenge the dual ontological status of what Popper called 'embodied' artifacts. The first argument is concerned with the material composition and characteristic features of artifacts. The second one addresses the creative and epistemic value of these artificial products.

2. Popper on the natural-artificial division and the partial autonomy of artifacts

The age-old philosophical debate about the natural and the artificial, the origin of which can be safely traced back to Aristotle, assumes that natural objects which exist, persist, have their nature and are classified independent of human beliefs, representations, experience, knowledge and practices are clearly different from artifacts which seem to depend for their existence, nature and classification on human beliefs, intentions, representations, knowledge and practices. Inverted, this assumption implies that artifacts do not figure among the 'furniture of the world' since they do not possess purely mind-independent discoverable natures. This apparent mind-dependence of artifacts continues to raise doubts about their *real* existence and the natural-artificial distinction is still a matter of intense dispute as can be witnessed in a series of articles published in *APA Newsletter on Philosophy and Computers* (2008).

2 By 'artifact' I understand any product of human intellectual and physical activities consciously conceived, manufactured or modified in response to some need, want or desire to produce an intended result. The term 'technical' is often used to differentiate artifacts designed mainly for practical purposes from aesthetic objects. For details see Kroes and Meijers (2006).

3See, for instance, the writings of Baker (2004, 2007, 2008), Thomasson (2003, 2008, 2009), Elder (1978, 2004), Verbeek (2005), Kroes and Vermaas (2008).

4By 'natural object' I mean that which is produced or developed by natural processes without minimal human intervention.

Lynne Baker (2008, pp.2-5), for instance, referring to the works of Wiggins (2001), questions the standard ways of singling out ontologically genuine substances and reasons that the mind-dependency of artifacts does not make them ontologically deficient as compared to natural objects. The alleged difference between natural objects and artifacts, she says rather pointedly, is steadily shrinking anyway because modern technology is creating products like digital organisms or bacterial batteries that are difficult to classify unambiguously as artifacts or natural objects. Beth Preston (2008, pp.26-28), on the contrary, argues that there never really was a sharp divide between natural objects and artifacts. Drawing attention to those ancient methods of domesticating plants or animals and primitive use of fermentation (which have nothing to do with advances in modern technology) she asserts that the natural-artificial divide was always blurry. On account of this she challenges the perceived significance of the more general distinction between mind-dependent and mind-independent objects that is often used to support the orthodox view of artifacts being ontologically deficient entities. Picking up on what is being debated for long Peter Kroes and Pieter Vermaas (2008, pp. 28-31) take Preston's side to argue that it is not due to modern technology that the difference appears fuzzy; it began to pose problems the moment human beings started using and modifying natural objects to meet their ends. But no matter how problematic this distinction appears, Kroes and Vermaas claim, there are some clear cut cases where the difference makes sense and is of great philosophical and pragmatic significance.

All these contemporary thinkers, regardless of their conflicting views, team up for challenging the almost unquestioned assumption underlying the natural-artificial distinction, namely, that objects existing, persisting, and being classified independent of human experience and knowledge are only to be considered as real. Baker has been highly applauded by others for breaking free of the traditional position and asserting that though artifacts depend on human minds or intentions in ways that natural objects do not, this does not imply any ontological deficiency in artifacts. Genuinely and objectively artifacts do qualify as real constituents of our world even if they are brought into being by our intellectual and physical activities, and in some sense 'up to us'. The appreciation of this crucial point that mind-dependence or intention-dependence[5] does not necessarily indicate ontological inferiority stimulates in turn the need to seek a broader image of reality that will enable us to grant artifacts a proper position in metaphysical schemes.

5 Here I assume no difference between mind-dependence and intention-dependence.

A possible solution to this appeal to a more comprehensive picture of reality can be found in Popper's theory (1972, 1977, 1979, 1982) of three ontologically distinct worlds, (namely, World 1, World 2, and World 3) acting upon and partially overlapping each other. This theory separates World 1 (the world of physical states, events, laws, animate and inanimate objects) from World 3 (the world of human creations, including artifacts) on the one hand and emphasizes the reality, objectivity, and partial autonomy of these World 3 products on the other. True, artifacts such as tools and machines do not hold center stage in Popper's exposition of the elements of World 3, seeing that theories, propositions, the abstract yet objective contents of scientific, mathematical or poetic thoughts, problem-situations and critical arguments are held by him as the 'most fertile' World 3 citizens (Popper 1972/1978, p.138). Nevertheless, this distinct world of human creation includes works of art, ethical values, social institutions and artifacts or what Popper (1979) calls, 'feats of engineering' such as, tools, machines, aircrafts, computers and scientific instruments as well. Drawing on the richness and diversity of the contents of this World 3, it would not be too difficult to extract an account of artifacts.

If artifacts are described as products of human minds, then on the face of it, they are mind-dependent entities. One can spot at least two different senses in which artifacts seem to be mind-dependent. The first sense of dependence is a simple causal matter; individual artifacts, such as, tables, chairs, books or computers are existentially dependent on human intentions as the intentional activities of humans are causally responsible for the creation of these entities. The other sense of mind-dependence is purely conceptual. Artifacts are conceptually dependent on human minds in the sense that it is 'metaphysically necessary' (Baker 2007) for something to be an artifact (as opposed to, say, a tree or a stone) that there be intentional human activities. Unlike garbage or pollution, artifacts, strictly speaking, are not *merely* the products of human activities, but the *intended* products of human activities (Hilpinen 1992, p.60). This very idea of mind-dependence of artifacts often makes metaphysicians hesitant to acknowledge their existence as it tends to suggest that human thought and intentions are sufficient to bring new entities into existence, like a rabbit in a hat by an element of magic or by a 'conjuring trick.' What lies behind the objections to artifacts on grounds of their (alleged) mind-dependence is basically this kind of worry. However, it does not in the least affect Popper for he never ever suggested that artifacts, like imaginary objects, can be brought into existence by human thought, intentions, beliefs or imagination *alone.* That the production of artifacts such as bridges or buildings essentially involves human physical activities was quite obvious to him (Popper 1972/1978, p.166). What is more important, this

mind-dependence of artifacts was in no way taken by Popper to interfere with their reality or (partial) autonomy. He neither doubted the reality (and partial autonomy) of these human products, nor did he ever hold them to be ontologically inferior to natural objects in any sense.

Possessing discoverable mind-independent natures (about which everyone may be turn out to be in error) is traditionally held to be the central criterion for treating entities as 'real' or genuine parts of our world (Schwartz 1978, Elder 1989). The implication is understandable: artifacts generally viewed as not having mind-independent natures accessible to scientific examination, are not real parts of the world. This traditional assumption often dubbed as the 'Aristotelian view of artifacts' has been challenged from two different perspectives. On the one hand, contemporary scholars advocate the necessity of questioning mind-independence as *the* criterion of real existence. For instance, Thomasson (2008, p. 25) argues, the very thought that to be real artifacts must have mind-independently discoverable natures is based on 'illegitimately generalizing from the case of scientific entities'. Hence this general, across-the board criterion of mind-independence as *the* criterion for the existence of 'anything whatsoever', she insists, should be given up. The criteria for existence may vary for different entities.

The other (relatively older) point of view (Simon 1969; Kornblith 1980; Losonsky 1990) upholds that although artifacts are our creations, they still may have intrinsic natures every bit as open to error or scientific discovery as the natures of chemical or biological kinds are. The popular proposal along these lines (Kornblith 1980) is that artifactual natures are at least largely distinguished by sameness of function rather than by sameness of chemical or genetic structure. But given the fragmentary nature of the existing philosophical accounts of artifact function[6] and the acknowledged limitations of this suggestion[7] I would like to limit myself to the submissions of Simon (1969) and Losonsky (1990).

Taking a closer look at one of the most extensively studied artifacts, namely, the clock, Simon (1969, pp.6-9) pointed out that the purposeful aspect of any artifact involves a relation among three terms, namely, the purpose or goal, the inner character of the artifact and the outer environment in which the artifact performs. The advantage of separating inner from outer environment in studying any artifact is that from knowledge of its purpose (or goal) and its

6 For a discussion on the philosophical theories of artifact function see Preston (2009).

7 Baker (2006) and Thomasson (2009) stand out among those who have been critical of this view.

outer environment its behavior can often be predicted. The clock will serve its intended purpose only if its 'inner environment' (say, for example, the arrangement of gears, the application of the forces of springs or gravity operating on a weight or pendulum) is appropriate to the 'outer' environment, the surroundings in which it is to be used. Sundials, for example, perform as clocks in sunny climates but are of no use at all during the Artic winter. Evidently, natural science impinges on an artifact through two these three terms of the relation that characterizes it: the inner structure of the artifact itself and the outer environment in which it will operate.[8]

Almost two decades later Losonsky (1990, pp.81-88) cited the same example of the clock to prove his point against the Aristotlelian view of artifacts. Artifacts, he argued, do have discoverable natures, and these natures underlie the changes artifacts undergo. One important feature of an artifact's nature is its internal form or structure, which contributes to its permanence and its reproduction. In addition to its inner structure, two more features, namely, the purposes for which it is used and how it is used for those purposes, also belong to its nature. Simply knowing how to use a clock, for instance, does not presuppose any familiarity with its internal nature. Since these three features (internal structure, purpose and manner of use) belong to the intrinsic nature of artifacts, Losonsky noted, it is no longer possible to believe that this nature is not worthy of scientific investigation or that this intrinsic nature does not underlie the ways in which artifacts develop and affect their environment.

In the circumstances, a careful scrutiny of Popper's pluralistic theory is worth-undertaking because it argues for a novel way of regarding artifacts as ontologically respectable aspects of reality without ignoring the fact of their mind-dependency and what is more, without entailing the requirement of having discoverable mind-independent natures. Two crucial claims regarding the ontological status of artifacts can be found in Popper much before they have been put forward by present-day philosophers. The claims are: first, artifacts being products of human creation are ontologically different from but not necessarily ontologically inferior to natural (that is, World 1) objects; second, the '*kickability*' of artifacts, that is, the fact that they can be kicked and can, in principle kick back (Popper 1982, p.116) is to be taken as evidence to substantiate their reality and (partial) autonomy. In what follows, I will examine these claims one by one.

8This division between inner and outer environments, can be found to a greater or lesser degree, in all large and complex systems, whether they are natural or artificial (Simon 1969, p.7).

Popper presented his thesis of three worlds against the then fashionable monistic materialism or the dualistic view of the universe. His argument for introducing an ontologically distinct World 3 rested primarily on the division he made between thought in the subjective sense (that is still a part of us) and thought in the objective sense (that is, thought formulated linguistically); in simple words, between World 2 *thought processes* and World 3 *thought contents*, a division neglected in traditional epistemology. Once formulated in language, any thought becomes an object outside ourselves and hence liable for inter-subjective criticism and evaluation. These objective contents of human thought possess various properties and relationships independently of any person's awareness of them. For instance, any scientific theory possesses (in a non-trivial sense) infinitely many logical consequences, yet the number of these consequences of which we can be aware of at any time is necessarily finite. Facts like this mean that the World 3 of objective contents must be distinguished both from World 2 (which consists of the various kinds of awareness we have of these objective contents) and from World 1(which consists of various forms of expressions of these objective contents) and therefore need to be classified into a separate class of things.

What makes any item an inmate of World 3, on Popper's view, is not as much the fact of its being a product of human creation as the fact that it can be grasped, known, deciphered or criticized inter-subjectively. Though originally generated by us, these World 3 objects, unlike ideas and thoughts (in the subjective World 2 sense), can be detached from the psychological processes of production and hence are potentially knowable, graspable, and analyzable. In other words, the very characteristic of World 3 objects is that they can be improved by cooperative criticism and criticism can come from people who had nothing to do with the original idea. The relevance of Popper's pluralistic thesis thus lies not only in his emphasis on the ontologically distinct character of these World 3 products but in his firm conviction that the question of the reality of these human creations can be addressed regardless of their psychological origin, or mind-dependency. This key Popperian insight exposes at once the insignificance of the mind-independence/mind-dependence question for the ontological status of any object. The wide-spread view that mind-dependency entails ontological deficiency, it is important to note, had been rejected by Popper decades before contemporary scholars wanted to get rid of it. What seems really at stake here is a problem that is of wider significance than the mind-(in) dependency issue, namely, the issue about the chief criterion for 'real' existence. This leads us straight into the other important claim put forward by Popper.

Something exists or is real, Popper taught us, if and only if, it can interact with members of World 1, with hard physical bodies. He (1979) took his cue from the physicalist's idea of reality. The physicalists more often than not were certain about the reality of medium sized physical objects that even a child can handle. Starting from this primitive idea of reality and then adopting the method of generalization they arrived at the idea of real physical existence by including very large and very small sized objects and also by including whatever can causally act upon things, such as magnetic and electrical attraction and repulsion, fields of forces as well as radiation, for example X-rays, because they can causally act upon bodies, say, photographic plates. Popper was thus led to the idea that what is real, is whatever, may directly or indirectly, have a causal effect upon physical bodies, and especially upon those physical bodies that can be easily handled. World 3 objects, he observed, do in fact strongly interact with the physical World 1 through the indispensable intervention of the subjective World 2 processes or the human mind.[9] Hence, the reality of those World 3 products is evident from the impact they make upon World 1 (via World 2), from their ability to have a profound feedback effect upon us by influencing our World 2 thought processes decisively, and from the impact any of us can make upon them. In short, the World 3 objects are real in the sense that they may have a causal effect upon our World 2 experiences, and further upon our brains belonging to World 1, and thus upon physical bodies.

The more noteworthy point regarding the contents of World 3 concerns their (partial) autonomous character. Once formulated in language or embodied materially these World 3 objects, *pace* Popper, begin to cause their own problems, to bring forth unintended, unforeseen consequences. In short, they express an autonomous aspect which is also real in the sense that it can interact with World 2 (as World 3 objects can have a strong causal influence upon our thought processes) and also, via World 2, with World 1. Popper's standard argument in support of this (partial) autonomy of World 3 comes in form of the following two thought experiments (Popper 1972/1978, pp.107-108; emphasis in original):

Experiment (1): All our machines and tools are destroyed, and all our subjective learning, including our subjective knowledge of machines and tools, and how to use them. But *libraries and our capacity to learn from them* survive. Clearly, after much suffering, our world may get going again.

9 In order that Special Relativity theory could have its influence upon the construction of the atom bomb, several physicists, Popper (1979) pointed out, have to get interested in the theory, work out its consequences, and *grasp* these consequences. This *grasping* or human *understanding* and thus the human mind seem to be quite indispensable.

Experiment (2): As before, machines and tools are destroyed, and all our subjective learning, including our subjective knowledge of machines and tools, and how to use them. But this time, *all libraries are destroyed also*, so that our capacity to learn from books becomes useless.

Popper conjectured that 'machines and tools,' in the absence of libraries, cannot help the reemergence of our civilization for many millennia. He seems optimistic about a civilization that has had its 'material infrastructure' destroyed, but still retains libraries and our 'capacity to learn from them'. Although it is not exactly clear from his writings quoted above whether his argument here is intended to devalue the (epistemic) importance of 'machines and tools' or whether it simply reflects his utter indifference to our capacity to learn from our experiences and uses of machinery, the epistemological merit of ideas, theories, or other linguistic products of human creation are noticeably more appreciated by him than the 'feats of engineering'.[10] One might here criticize Popper for failing to recognize the epistemic importance of the material products of human manufacture, but his point on the (partial) autonomy of World 3 products (including artifacts) deserves attention.

The notion of 'autonomy' seems to be a problematic one and philosophers concerned with technology and technological products, are arguing over this concept for quite some time. Tenner's (2006) classification of an enormous number of technologies which end up having disastrous or unpredictable consequences is quite well known. Such examples lead right to the question: do technical artifacts have a life of its own? Drawing on the old Greek idea that artificiality implies controllability, Pitt (2011, pp.73-83) reasons that for technology to be autonomous, it must be uncontrollable. Since we do control, challenge, change, and even reject technology including the large-scale ones (though not all of it, not all the time), the very question of technology being autonomous is not to be entertained.

Popper's idea of autonomy, however, appears very different from what Pitt and others understand by this term. Artifacts (and all other World 3 contents) despite being products of the workings of innumerable minds do have a life more independent of human intention and endeavor as they bring forth unintended, unforeseen consequences. It is in this sense, Popper understood, they are to a considerable extent autonomous. Unfortunately the examples discussed by Popper are taken mostly from mathematics and except for a few comments on the impact of nuclear reactors or atom bombs on humanity he did not ponder much on the

10 In his Tanner Lectures (Popper 1979) he admitted openly that scientific conjectures can exert a much stronger effect (via World 2) upon physical things than technical artifacts such as scissors and screwdrivers.

autonomous character of artifacts. Nevertheless, the real significance of his argument in defense of the (partial) autonomy of World 3 creations comes to light as soon as one reflects on the nature of our dynamic relationships with artifacts. A closer look into Ihde's (1979) phenomenological analysis of how technical artifacts 'mediate' human-world relations seems most suitable for understanding Popper's notion of autonomy.

One of the most interesting examples provided by Ihde (1979, pp.18-23) is that of a dentist using her probe to gather information about our teeth. Certain features of the dentist's experience are to be noted. The finely tipped probe exists 'between' the dentist and what is experienced and in this sense is the 'means' of her experience of the texture, hardness, softness, or holes of our tooth. The dentist feels the hardness or softness 'at the end of the probe'. She discerns that as she experiences the tooth through the probe, the probe is being taken into her 'self-experiencing'. This has an interesting implication, namely, that here touch is 'at a distance', and touch at a distance calls for some material embodiment. However, one also needs to note the converse side of the sense of touch at a distance. Simultaneous to the awareness of the tooth as the focal object of her experience, there is the 'relative disappearance' of the probe as such.[11]

This disappearance or withdrawal is the way the instrument becomes the 'means' by which 'I' can be extended beyond my bodily limit. It may thus be spoken of as a withdrawal into my now extended 'self-experience'. The probe genuinely extends the dentist's awareness of the world, it allows her to be embodied at a distance, and it amplifies certain characteristics of the tooth as well. It gives her what, compared to 'in the flesh' experience, are micro- features of the tooth's surface. But at the same time that the probe extends and amplifies, it reduces another dimension of the tooth experience. With her finger the dentist can sense the warmth or wetness of the tooth, aspects which she does not get through the probe at all. The probe, precisely in giving her a finer discrimination related to the micro-features, reduced the full range of other features sensed in her finger's touch. The dentist experiences the tooth through the probe, but it is equally clear that what is experienced is in some ways transformed and quite different from 'in the flesh' experiences.

We just saw how a simple stainless steel probe transforms direct perceptual experience. Artifacts, therefore, are not 'neutral intermediaries' between humans and world, but

11 It was Heidegger (1927/1962) who first observed this peculiarity of what is proximally ready-to-hand. He pointed out that in order to be ready-to-hand it must quite authentically withdraw (*zurueckzuziehen*).

'mediators'; they actively 'mediate' this relation.[12] This, what Ihde calls, 'non-neutrality' of artifacts can be seen as expressive of what Popper refers to as their (partial) autonomy. Though artifacts are our products, creations of our intellectual and physical efforts, they are to a large extent autonomous in this particular sense that they have the potential to transform our experience, to affect our actions, our everyday dealings with the world, in unanticipated or unintended ways. As they become part of our self-experience and self-expression we, Popper (1972/78, pp.146-150) felt, are able to transcend ourselves (that is, our talents, our gifts) through our dynamic and incessant interaction with our own creations. Probably because of our obsession with representation and theory at the expense of action and intervention that such dynamic autonomous character of artifacts is scarcely noticed in mainstream philosophical discussions.

3. Popper on the ontological status of artifacts

Popper (1977/1995, 1982) drew an interesting distinction between 'embodied' and 'un-embodied' World 3 objects that is, between products of human mind that are linguistically formulated or materially constituted and those that are not yet so constituted or formulated. An un-embodied World 3 product, for instance, may be any hitherto unexplored logical problem situation, or hitherto undiscovered logical relations between existing theories. This distinction between embodied and un-embodied World 3 products is not to be confused with the general division of artifacts into categories of 'material' and 'abstract.' Dasgupta (1996, pp. 9-12) classifies architectures, plans, designs, etc. (which are rendered visible through symbol structures) as abstract artifacts, because though they are artificial products intended to serve certain human purposes, they are materially intangible in form.[13] Important to note, while the architectural plan of a building (symbolically formulated) is an 'abstract' artifact for Dasgupta, Popper classified it as an 'embodied' World 3 product.

Some embodied objects like books, paintings, or sculptures, Popper argued, have a dual (ontological) status. Let us consider his favorite example of a book. As a tangible physical entity it belongs to World 1, but in so far as it has a content that remains invariant through various editions and can be examined for matters like logical consistency, it belongs simultaneously to World 3. Similarly, sculptures, paintings etc. being receptacles of objective

12Not all experiences with artifacts, however, are of this type. For a detailed view see Ihde (1979).

13 The way one can touch a material artifact say, a building, the architectural plan of it cannot be touched in the same way.

content are inmates both of World 1 and World 3. Dasgupta (1996), Eccles (1974) and Baird (2004) stand out among those for whom this pluralistic (Popperain) thesis advanced to challenge the traditional Cartesian categorization of the universe into objective physical reality and subjective mental events holds great promise. However, neither of them approves this dual-status of embodied objects. Whereas Dasgupta (1996) and Eccles (1974) place materially constituted artifacts directly in World 1, Baird (2004) suggests that material artifacts, though not linguistically built, should belong exclusively to World 3.

Until and unless one could spell out what difference there is, if any, between regular World 1 objects and those material structures which being possessors of objective contents of thought belong simultaneously to World 3, this proposal of the dual-status of embodied World 3 products seems to leave a lot to be desired. In what follows, I try to offer two arguments to question this dual ontological status of embodied artifacts and to reinforce Baird's (2004) suggestion that artifacts should belong exclusively to World 3, a distinct world of human creation. The views of Eccles (1974) and Dasgupta (1996) regarding the (ontological) categorization of material artifacts in World 1 are rejected by implication.

First and foremost, I would like to argue, that artifacts, despite their physical-chemical make-up cannot, strictly speaking, be inhabitants of World 1 since the internal substance and organization of any artifact (materially constituted), in contrast to a natural object (in the sense clarified in footnote 4) is an 'engineered' or 'designed' structure that bears clear traits of human involvement[14] and not simply a given assemblage of raw materials. The components of any material artifact, say a pencil, are not 'raw' in the sense that naturally occurring materials like clay or wood are raw, rather they are skillfully and carefully selected, organized, modified, processed or in part refurbished, demonstrating signs of human interference all over. To cite another example, though a rubber ball is immediately made of rubber, it is not to be identified with the part of rubber of which it is composed. That part of rubber may have been synthesized before being formed into a spherical shape to create the ball, and certainly the part of rubber could continue to exist (in some sense) even if the ball were to be destroyed.[15] As this inner (physical-chemical) structure of any material artifact, in virtue of which it is generally thought to belong to World 1, is an engineered or designed structure, artifacts, it seems safe to hold, are clearly different from natural objects and do not belong to the 'given', natural World 1.

14Even the pre-historic stone tools (axes, hammers etc.) were made by chipping and flaking techniques that required skilled human labor.

More notably, artifacts are generally characterized by a certain 'for-ness', that is, they have a functional or purposeful aspect.[16] However, though they are products designed for human purposes, their purposeful or functional nature is neither wholly determined by the physical properties of the constituents nor by external physical factors (such as physical laws or forces) and also cannot be explained in complete isolation from the socio-cultural context of their use.[17] In short, the fulfillment of purpose or the realization of function does not wholly depend on the inner physical structure of the artifact in any important sense. The main reason being, artifact functions are typically *multiply realizable*, that is, they are realizable in a variety of materials and/or forms, provided some general constraints are satisfied. As Preston (2009) illustrates, spoons have to be made out of a relatively rigid material and have a form that includes a handle attached to a bowl. But other than that form and material are very variable. Since a given artifact function is realizable in a range of forms and materials, it is no wonder that it can also be performed by other artifacts originally designed to fulfill different functions. Therefore artifacts are *multiply utilizable*; typically they serve several functions, often simultaneously. For example, an umbrella designed specifically to ward off rain or to be used as a sunshade, can also be used as a weapon, as a lampshade, as a handy extension of the arm for reaching and retrieving things.[18] Hence the mere possession of a tangible structure or certain physical-chemical-geometrical properties cannot be a sufficient ground for placing

15I do not raise the problem of coinciding objects here for the following reason. The most popular view often referred to as the 'standard account' (Lowe 1995) embraces the conclusion that numerically distinct objects, (for instance, a certain wooden table and the lump of wood which composes it) can exist in the same place at the same time. The underlying assumption is: all that needs to be done to a lump of wood in order to make it into a table is to merely change its shape in an appropriate way. Considering contemporary philosophical and engineering research on the design and manufacture of artifacts (Bucciarelli 1994; Vermaas et al. 2008) I find this assumption too simple to go entirely unquestioned.

16 It has recently been argued that technical artifacts have a dual-nature: they are designed physical structures, which realize functions that refer to human intentionality (Kroes and Meijers 2006). Artifact functions are commonly believed to be directly and exhaustively determined by individual and/or collective human intentions. Though there are scholars (Preston 2009) who doubt whether functions of artifacts are dependent on human intentions in any relevant sense, at least this much is admitted on all hands that human intentions have *something* to do with the functions of artifacts.

17 For a detailed view on the relevance of socio-cultural factors see, for instance, Basalla (1988), Preston (2006), Priemus & Kroes (2008).

18 No doubt artifacts have standardized forms and uses that are (relatively) stable for years or even generations. What needs to be emphasized is that they are only *relatively* stable.

artifacts in World 1. Compositionally and characteristically they differ from natural objects, the inmates of World 1.

Before presenting the second argument it is important to recall the Popperain notion of objective knowledge which consists of ideas, problems, theories, arguments – coded symbolically in the actual material structures serving as vehicles for this knowledge so that their objective existence in ensured and in fact can continue independently of anybody's claim to know them or know about them. Popper's pluralistic thesis implies an ontological division between the 'material structure' of an artifact and the 'objective content or knowledge' that this structure is a carrier of. For example, the material structure of a book made out of paper, glue, thread etc. is ontologically distinct from its abstract content possessing certain semantic and syntactic properties. This division clearly rests on the assumption that the three-dimensional material structure is simply a carrier of objective content or knowledge and hence cannot be a part of World 3. Two reasons can be offered to contest this underlying assumption.

First of all, Popper seems to overlook the fact that the material structure is as much a product of creative imagination, rational thinking and inter-subjective criticism as the content it embodies. The act of conceptualizing and manufacturing the structural forms of artifacts intended to meet given human requirements is technically known as design. Design is typically conceived of as a purposeful, goal-directed activity, a process of making something that has never existed before. Such a task-specific process would only be initiated if there is no existing artifact that perfectly fulfills the given requirements. As novelty or originality, even in the most modest sense, is a condition needed for the process of design to begin, the design-process is widely viewed as a creative process.[19] In saying this I do not mean to endorse the traditional hylomorphic model of creation which entails the idea of form (*morphe*) to be imposed by an agent with a specific goal in mind on passive and inert matter (*hyle*). I am quite aware that in contemporary discussions in fields ranging from artifact-design (Franssen 2008; Ihde 2008) to material culture studies (Ingold 2007) a tendency to counteract this widespread view is already visible. Designers are no longer seen as having a great deal of control over the design-process and the roles played by historical choices,

19This, however, is not to suggest that every act of design counts as a creative act in the most elevated sense of the term. A closer look into Dasgupta's (1996, pp.53-65) analysis of different levels of creativity would be very helpful at this point.

cultural assumptions and social contingencies in the creative process of artifact-design are being seriously considered.

On the other hand, it is presently argued (Ingold 2007) that the material world is not passively subservient to human designs; the forms of things cannot be imposed from without upon an inert substrate of matter as matter is always in flux, in variation. In the generation of things the materials with various and variable properties enlivened by the forces of the cosmos actually meld with one another. Therefore, the creativity of the work is to be seen in the forward movement that gives rise to things, in joining with and following the forces and flows of materials as they unfold and bring the novel form into being. Here the processes of genesis and growth that bring about forms in the world are viewed as more important than the finished forms themselves.

Whether one should assign primacy to processes of formation as against their final products is too big a question to be discussed at this point. In the hylomorphic model of creation creativity is to be read backwards, starting from an outcome in the form of a novel object and tracing it, through a sequence of antecedent conditions, to an unprecedented idea in the mind of an agent or designer. The new alternative, on the contrary, puts more emphasis on the processes of form-giving than on the finished forms themselves and spots creativity in this forward movement that generates things. Irrespective of the view one chooses to hold up, the fact remains that material structures or forms of artifacts brought forth by the processes of design are products of human ingenuity and not elements of the 'given' physical world. Hence they should belong to World 3, the world of human creation.

The second reason concerns the epistemological merit of materially constituted artifacts completely neglected by Popper. The material form or structure of any artifact (say a book or a microscope), it has recently been shown (Baird 2004), is not only instrumental to the articulation of knowledge expressed in words but is a specimen or token of knowledge[20] itself. Although this idea of 'thing-knowledge' has been explicitly pointed out by Baird (2004) lately, the germ of this idea that technical devices, their pictures and drawings can convey a vast body of characteristically non-verbal knowledge can be traced back to Ferguson (1977). One might also be tempted to ask how material artifacts could represent knowledge when, as we are accustomed to believe, knowledge requires semantic content and

20 The term 'knowledge' is used here in the objective sense as discussed by Popper (1972). In the objective sense knowledge can be understood as an evolutionary product of human (intellectual and physical) activities that can be detached from its psychological origin, can be criticized and modified inter-subjectively, and can improve our active adaptation to the world.

hence must be propositional in nature. In sharp contrast to our traditional attitude of thinking about knowledge in propositional terms, and of considering theories as the primary means for expressing knowledge, Baird (2004) advances a 'materialist epistemology'. This materialist epistemology focuses on technical artifacts (like instruments for scientific experiments, observation or measurement) not simply because of their role in the generation, articulation or justification of knowledge (expressed linguistically) but because they bear knowledge themselves, on a par with the words we speak and hence are epistemologically valuable in their own right. The knowledge borne by things is typically different from knowledge that our theories bear, and cannot obviously be described as 'justified true belief'. Baird (2004) discusses three different kinds of knowledge, namely, model knowledge, working knowledge, and encapsulated knowledge, borne by scientific instruments in order to demonstrate that they do have epistemic content and understanding that content is important to a more comprehensive account of science.

While Baird considers mainly scientific instruments like Faraday's first electric motor, and direct reading spectrometers, etc. to illustrate his thesis, I intend to suggest that not only high-profile scientific instruments but such seemingly simple everyday artifacts like pins and paperclips are instances of knowledge too. Each artifact itself is a unique manifestation of human imagination, workmanship and of quite a rich combination of knowledge. The knowledge embodied by these material artifacts is notably heterogeneous in nature. It may include, formal engineering knowledge (generally called technological theory), mathematics, knowledge of the sciences, theoretical tools (e.g. calculation methods for forces in a construction), and most importantly what Polanyi (1962) called knowledge of 'operational principles'[21] that often remains tacit. Drawing on Petroski's (1992) painstaking research on the evolution of everyday artifacts I try to indicate in what way a simple and mundane paper clip can be seen as a (non-verbal) expression of knowledge and as epistemologically important in its own right.

A paper clip (successfully working) is usually made with a steel wire that wants to spring back to its original shape after being bent, but only up to a point, for otherwise the paper clip could not be formed into the object it is. The paper clip works because its loops can be spread apart just enough to get it around some papers and, when released, can spring back to grab

21Inspired by Polanyi, Dasgupta (1996, p.158) defines an operational principle as any proposition, rule, procedure, or conceptual frame of reference about artifactual properties that facilitate action for the creation, manipulation, and modification of artifactual forms and their implementation.

the papers and hold them tight. This springing action, more than its shape per se, is what makes the paper clip work.[22] Robert Hooke discovered the nature of this spring force in 1660 and published his observation about the elasticity or springiness of materials in 1668. There must be the 'right spring' to the paper clip wire, and to try to make clips with too stiff or too soft a wire is tantamount to trying to break Hooke's Law.[23] A paper clip then encapsulates in its material form the knowledge of the characteristic springiness of materials and the knowledge of how to apply the 'right spring' to the paper clip wire. The former is the scientific knowledge of the fundamental behavior of materials, while the latter is an operational principle. As an instance or non-verbal expression of (objective) knowledge itself the paperclip should reasonably belong to World 3. This seems to hold true for other materially constituted artifacts as well. The Popperian suggestion of the dual ontological status of embodied World 3 products thus needs to be dropped.

Since artifacts too like ideas and theories are (non-verbal) expressions of knowledge, the traditional questions of the character and growth of knowledge need to be reconfigured in the light of new questions concerning the things we make. For instance, to consider technical artifacts as instances of knowledge amounts to questioning the basic postulation of the traditional philosophical theory (of knowledge), namely, that knowledge consists of those beliefs which can be justified. In addition this involves a rethinking of the notions of truth and justification which are tied to the concept of knowledge but seem hard to fit around artifacts.[24] It is high time philosophers of science and technology ought to be concerned with the ways human knowledge is embedded in such technological products.

REFERENCES

1. Baker, L. R. 2004. The Ontology of Artifacts. *Philosophical Explorations* 07: 99-112.

22 Every material that engineers work with, whether it is timber, iron, or steel wire has a characteristic springiness to it.

23 That means if one were to use wire that did not stay bent, then the loop could not even be formed. On the other hand, if one were to try to make a paper clip out of wire that stayed bent too easily, it would have little spring and not hold papers very tightly.

24 For an interesting discussion of these issues see Pitt (1998) and Baird (2004).

2. Baker, L. R. 2007. *The Metaphysics of Everyday Life: An Essay in Practical Realism.* Cambridge: Cambridge University Press.

3. Baker, L. R. 2008. The Shrinking Difference between Artifacts and Natural Objects. *APA Newsletter on Philosophy and Computers* 07 (2): 2-5.

4. Basalla, G. 1988. *The Evolution of Technology*. Cambridge: Cambridge University Press.

5. Baird, D. 2004. *Thing Knowledge: A Philosophy of Scientific Instruments*. Berkeley: University of California Press.

6. Bucciarelli, L. L. 1994. *Designing Engineers*. Cambridge, MA: MIT Press.

7. Carr, B. 1977. Popper's Third World. *The Philosophical Quarterly* 27: 214-226.

8. Cohen, L. J. 1980. Some Comments on Third World Epistemology. *The British Journal of Philosophy of Science* 31: 175-180.

9. Currie, G. 1978. Popper's Evolutionary Epistemology. *Synthese* 37: 413-431

10. Dasgupta, S. 1996. *Technology and Creativity*. Oxford, New York: Oxford University Press.

11. Dipert, R. 1993. *Artifacts, Artworks, and Agency.* Philadelphia, PA: Temple University Press.

12. Eccles, J. C. 1974. The World of Objective Knowledge. In P. A. Schilpp (Ed.), *The Philosophy of Karl Popper*. Illinois: The Open Court – La Salle

13. Elder, C. 1989. Realism, Naturalism and Culturally Generated Kinds. *Philosophical Quarterly* 39: 425-444.

14. Elder, C. 2004. *Real Natures and Familiar Objects*. Cambridge, MA: MIT Press

15. Ferguson, E. S. 1977. The Mind's Eye: Nonverbal Thought in Technology. *Science* 197(4306): 827-836.

16. Franssen, M. 2008. Design, Use, and the Physical and Intentional Aspects of Technical Artifacts. In P.E. Vermaas et al. (Eds.), *Philosophy and Design.* Springer.

17. Hacking, I. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press

18. Heidegger, M. 1927/1962. *Being and Time*, Translated by J.Macquarrie and E.Robinson Oxford: Basil Blackwell

19. Hilpinen, R. 1992. Artifacts and Works of Art. *Theoria* 58: 58-82

20. Ihde, D. 1979. *Technics and Praxis.*Dordrecht: D. Reidel.

21. Ihde, D. 1990. *Technology and the Lifeworld.* Bloomington: Indiana University Press.

22. Ihde, D. 1991. *Instrumental Realism*. Bloomington: Indiana University Press

23. Ihde, D. 2008. The Designer Fallacy and Technological Imagination. In P.E. Vermaas et al. (Eds.), *Philosophy and Design.* Springer.

24. Ingold, I. 2007. Materials against Materiality. *Archaeological Dialogues* 14(1): 1-16.

25. Kornblith, H. 1980. Referring to Artifacts. *The Philosophical Review* 89: 109-114.

26. Kroes, P. and Meijers, A.W.M. 2006. The Dual-Nature of Technical Artifacts. *Studies in History and Philosophy of Science* 37(1): 1-4.

27. Kroes, P. and Vermaas, P.E. 2008. “Interesting Differences between Artifacts and Natural Objects”. *APA Newsletter on Philosophy and Computers* 08 (1): 28-31.

28. Losonsky, M. 1990. The Nature of Artifacts. *Philosophy* 65 (251): 81-88.

29. Lowe, E. J. 1995. Coinciding Objects: In Defense of the ‘Standard Account.’ *Analysis* 55(03):171-178.

30. Petroski, H. 1992/1994. *The Evolution of Useful Things*. New York: Vintage Books.

31. Pitt, J. C. 2011. *Doing Philosophy of Technology*. Dordrecht: Springer.

32. Polanyi, M. 1962. *Personal Knowledge*. London: Routledge & Kegan Paul.

33. Popper, K. R. 1972/1978. *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.

34. Popper, K. R. and Eccles, J. C. 1977/1995. *The Self and Its Brain: An Argument for Interactionism*. London: Routledge.

35. Popper, K. R. 1979. Three Worlds. *Michigan Quarterly Review* 18 (1): 1-23.

36. Popper, K. R. 1982. *The Open Universe: An Argument for Indeterminism*. London: Hutchinson & Co. Ltd.

37. Preston, B. 2006. Social Context and Artifact Function. *Studies in History and Philosophy of Science* 37: 37- 41.

38. Preston, B. 2008. The Shrinkage Factor: Comment on Lynne Rudder Baker's "The Shrinking Difference between Artifacts and Natural Objects". *APA Newsletter on Philosophy and Computers* 08 (1): 26-28.

39. Preston, B. 2009. Philosophical Theories of Artifact Function. In A. W. M. Meijers (Ed.), *Philosophy of Technology and Engineering Sciences.* Elsevier Science.

40. Priemus, H. and Kroes, P. 2008. Technical Artifacts as Physical and Social Constructions: The Case of Cite` de la Muette. *Housing Studies* 23 (5): 717-736.

41. Schwartz, S. 1978. Putnam on Artifacts. *Philosophical Review* 87(04): 566-574.

42. Simon, H. A. 1969. *The Sciences of the Artificial.* MIT: The Riverside Press.

43. Tenner, E. 1996. *Why Things Bite Back: Technology and the Revenge of Unintended Consequences*. New York: Alfred A. Knopf.

44. Thomasson, A. L. 2003. Realism and Human Kinds. *Philosophy and Phenomenological Research* 67(03): 580-609.

45. Thomasson, A. L. 2008. Artifacts and Mind-Dependence: Comments on Lynne Rudder Baker's 'The Shrinking Difference between Artifacts and Natural Objects. *APA Newsletter on Philosophy and Computers* 08 (1): 25-26.

46. Thomasson, A. L. 2009. Artifacts in Metaphysics. In A.W.M. Meijers (Ed.), *Philosophy of Technology and Engineering Sciences.* Elsevier Science.

47. Van Inwagen. 1990. *Material Beings*. Ithaca: Cornell University Press.

48. Verbeek, P. P. 2005. *What Things Do:Philosophical Reflections on Technology, Agency, and Design*. University Park PA: Penn State University Press.

49. Wiggins, D. 2001. *Sameness and Substance Renewed.* Cambridge: Cambridge University Press.

On the Debate Concerning the Proper Characterisation of Quantum Dynamical Evolution

Michael E. Cuffaro[†] and Wayne C. Myrvold[*]

[†,*]The University of Western Ontario, Department of Philosophy

June 17, 2012

Abstract

There has been a long-standing and sometimes passionate debate between physicists over whether a dynamical framework for quantum systems should incorporate not completely positive (NCP) maps in addition to completely positive (CP) maps. Despite the reasonableness of the arguments for complete positivity, we argue that NCP maps should be allowed, with a qualification: these should be understood, not as reflecting 'not completely positive' evolution, but as linear extensions, to a system's entire state space, of CP maps that are only partially defined. Beyond the domain of definition of a partial-CP map, we argue, much may be permitted.

1 Introduction

Conventional wisdom has it that any evolution of a quantum system can be represented by a family of completely positive (CP) maps on its state space. Moreover, there seem to be good arguments that evolutions outside this class must be regarded as unphysical. But orthodoxy is not without dissent; several authors have argued for considering evolutions represented by maps that are not completely positive (NCP).

The debate has implications that have the potential to go deep. The possibility of incorporating NCP maps into our quantum dynamical framework may illuminate much regarding the nature of and relation between quantum entanglement and other types of quantum correlations (Devi et al., 2011). If the use of NCP maps is illegitimate however, such investigations must be dismissed without further ado.

In the following, we will argue for the proposition that NCP maps should be allowed—but we will add a caveat: one should not regard NCP dynamical maps as descriptions of the 'not completely positive evolution' of quantum systems. An 'NCP map', properly understood, is a linear extension, to a system's entire state space, of a CP map that is only defined on a subset of this state space. In fact, as we will see, not much constrains the extension of a partially defined CP map. Depending on the characteristics of the state preparation, such extensions may be not completely positive, inconsistent,[1] or even nonlinear.

The paper will proceed as follows: in Section 2 we review the essential aspects of the theory of open quantum systems and in Section 3 we present the standard argument for complete positivity. In Section 4 we consider the issues involved in the debate over NCP maps and in

[1]Strictly speaking, when an inconsistent map is used this should not be seen as an extension but as a change of state space. This will be clarified below.

Section 5 we present our interpretation of the debate and what we believe to be its resolution.

2 Evolution of a Quantum System

Consider a quantum system S that is initially in a state ρ_S^0, represented by a density operator $\hat{\rho}_S^0$. If the system is isolated, its evolution will be given by a one-parameter family of unitary operators $\{U^t\}$, via

$$\hat{\rho}_S^t = U^t \, \hat{\rho}_S^0 \, U^{\dagger t}. \tag{1}$$

Suppose, now, that the system interacts with another system R, which may include some piece of experimental apparatus. We take R to include everything with which S interacts. Suppose that S is prepared in a state that is uncorrelated with the state of R (though it may be entangled with some other system, with which it doesn't interact), so that the initial state of the composite system $S + R$ is

$$\hat{\rho}_{SR}^0 = \hat{\rho}_S^0 \otimes \hat{\rho}_R^0. \tag{2}$$

The composite system will evolve unitarily:

$$\hat{\rho}_{SR}^t = U^t \, \hat{\rho}_{SR}^0 \, U^{\dagger t}, \tag{3}$$

where now $\{U^t\}$ is a family of operators operating on the Hilbert space $\mathcal{H}_S \otimes \mathcal{H}_R$ of the composite system. It is easy to show (see, e.g., Nielsen and Chuang 2000, §8.2.3) that, for each t, there will be a set $\{W_i(t)\}$ of operators, which depend on the evolution operators $\{U^t\}$

and the initial state of R, such that

$$\hat{\rho}_S^t = \sum_i W_i(t)\, \hat{\rho}_S^0\, W_i^\dagger(t);$$
$$\sum_i W_i^\dagger(t) W_i(t) = I. \tag{4}$$

This is all in the Schrödinger picture, in which we represent a change of state by a change in the density operator used. We can also use the Heisenberg picture, which represents a state change via a transformation of the algebra of operators used to represent observables:

$$\rho_S^t(A) = \rho_S^0(A^t), \tag{5}$$

where

$$A^t = \sum_i W_i(t) A^0 W_i^\dagger(t). \tag{6}$$

In addition to unitary evolution of an undisturbed system, we also associate state changes with measurements, via the collapse postulate. In the case of a von Neumann measurement, there is a complete set $\{P_i\}$ of projections onto the eigenspaces of the observable measured, and the state undergoes one of the state transitions T_i given by

$$\mathcal{T}_i\hat{\rho} = \frac{P_i\, \hat{\rho}\, Pi}{\mathrm{Tr}(P_i\, \hat{\rho})}, \tag{7}$$

The probability that the state transition will be $\mathcal{T}_i$ is $\mathrm{Tr}(P_i\, \hat{\rho})$. When a measurement has been performed, and we don't yet know the result, the state that represents our state of knowledge of the system is

$$\mathcal{T}\hat{\rho} = \sum_i P_i\, \hat{\rho}\, P_i. \tag{8}$$

Note that this, also, has the form (4).

One can also consider *selective* operations, that is, operations that take as input a state and yield a transformed state, not with certainty, but with some probability less than one, and fail, otherwise. One such operation is the procedure of performing a measurement and keeping the result only if the outcome lies in a specified set (for example, we could do a spin measurement and select only '+' outcomes); the operation fails (does not count as preparing a state at all) if the measurement yields some other result. A selective operation is represented by a transformation of the state space that does not preserve norm. A selective operation $\mathcal{T}$, applied to state ρ, produces a final state $\mathcal{T}\rho$ with probability $\mathcal{T}\rho(I)$, and no result otherwise.

Unitary evolution, evolution of a system interacting with an environment with which it is initially correlated, and measurement-induced collapse can all be represented in the form (4). The class of state transformations that can be represented in this form is precisely the class of *completely positive* transformations of the system's state space, to be discussed in the next section.

3 Completely Positive Maps

We will want to consider, not just transformations of a single system's state space, but also mappings from one state space to another. The operation of forming a reduced state by tracing out the degrees of freedom of a subsystem is one such mapping; as we will see below, assignment maps used in the theory of open systems are another.

We associate with any quantum system a C^*-algebra whose self-adjoint elements represent the observables of the system. For any C^*-algebra $\mathcal{A}$, let $\mathcal{A}^*$ be its dual space, that is, the set of bounded linear functionals on $\mathcal{A}$. The state space of $\mathcal{A}$, $\mathcal{K}(\mathcal{A})$, is the subset of $\mathcal{A}^*$ consisting of positive linear functionals of unit norm.

For any linear mapping $\mathcal{T} : \mathcal{A} \to \mathcal{B}$, there is a dual map $\mathcal{T}^* : \mathcal{A}^* \to \mathcal{B}^*$, defined by

$$\mathcal{T}^*\mu(A) = \rho(\mathcal{T}A) \text{ for all } A \in \mathcal{A}. \tag{9}$$

If $\mathcal{T}$ is positive and unital, then $\mathcal{T}^*$ maps states on $\mathcal{A}$ to states on $\mathcal{B}$. Similarly, for any mapping of the state space of one algebra into the state space of another, there is a corresponding dual map on the algebras.

For any n, let W_n be an n-state system that doesn't interact with our system S, though it may be entangled with S. Given a transformation $\mathcal{T}$ of the state space of S, with associated transformation $\mathcal{T}$ of S's algebra, we can extend this transformation to one on the state space of the composite system $S + W_n$, by stipulating that the transformation act trivially on observables of W_n.

$$(\mathcal{T}^* \otimes I_n)\rho(A \otimes B) = \rho(\mathcal{T}(A) \otimes B). \tag{10}$$

A mapping $\mathcal{T}^*$ is *n-positive* if $\mathcal{T}^* \otimes I_n$ is positive, and *completely positive* if it is n-positive for all n. If S is a k-state system, a transformation of S's state space is completely positive if it is k-positive.

It can be shown (Nielsen and Chuang, 2000, §8.2.4) that, for any completely positive map $\mathcal{T}^* : \mathcal{K}(\mathcal{A}) \to \mathcal{K}(\mathcal{B})$, there are operators $W_i : \mathcal{H}_\mathcal{A} \to \mathcal{H}_\mathcal{B}$ such that

$$\begin{gathered}\mathcal{T}^*\rho(A) = \rho(\textstyle\sum_i W_i^\dagger \, A \, W_i); \\ \textstyle\sum_i W_i^\dagger W_i \leq I.\end{gathered} \tag{11}$$

This is equivalent to a transformation of density operators representing the states,

$$\hat{\rho} \rightarrow \hat{\rho}' = \sum_i W_i \, \hat{\rho} \, W_i^\dagger. \tag{12}$$

The standard argument that any physically realisable operation on the state of a system S must be completely positive goes as follows. We should be able to apply the operation $\mathcal{T}^*$ to S regardless of its initial state, and the effect on the state of S will be the same whether or not S is entangled with a "witness" system W_n. Since S does not interact with the witness, applying operation T^* to S is equivalent to applying $\mathcal{T}^* \otimes I_n$ to the composite system $S + W_n$. Thus, we require each mapping $\mathcal{T}^* \otimes I_n$ to be a positive mapping, and this is equivalent to the requirement that $\mathcal{T}^*$ be completely positive.

To see what goes wrong if the transformation applied to S is positive but not completely positive, consider the simplest case, in which S is a qubit. Suppose that we could apply a transformation $\rho_S^0 \rightarrow \rho_S^1$ that left the expectation values of σ_x and σ_y unchanged, while flipping the sign of the expectation value of σ_z.

$$\rho_S^1(\sigma_x) = \rho_S^0(\sigma_x); \quad \rho_S^1(\sigma_y) = \rho_S^0(\sigma_y); \quad \rho_S^1(\sigma_z) = -\rho_S^0(\sigma_z). \tag{13}$$

Suppose that S is initially entangled with another qubit, in, e.g., the singlet state, so that

$$\rho_{SW}^0(\sigma_x \otimes \sigma_x) = \rho_{SW}^0(\sigma_y \otimes \sigma_y) = \rho_{SW}^0(\sigma_z \otimes \sigma_z) = -1. \tag{14}$$

If we could apply the transformation (13) to S when it is initially in a singlet state with W,

this would result in a state ρ^1_{SW} of $S+W$ satisfying,

$$\rho^1_{SW}(\sigma_x \otimes \sigma_x) = \rho^1_{SW}(\sigma_y \otimes \sigma_y) = -1; \quad \rho^1_{SW}(\sigma_z \otimes \sigma_z) = +1. \tag{15}$$

This is disastrous. Suppose we do a Bell-state measurement. One of the possible outcomes is the state $|\Psi^+\rangle$, and the projection onto this state is

$$|\Psi^+\rangle\langle\Psi^+| = \frac{1}{4}\left(I + \sigma_x \otimes \sigma_x + \sigma_y \otimes \sigma_y - \sigma_z \otimes \sigma_z\right). \tag{16}$$

A state satisfying (15) would assign an expectation value of $-1/2$ to this projection operator, rendering it impossible to interpret this expectation value as the probability of a Bell-state measurement resulting in $|\Psi^+\rangle$.

Note that the set-up envisaged in the argument is one in which it is presumed that we can prepare the system S in a state that is uncorrelated with the active part of its environment R. This set-up includes the typical laboratory set-up, in which system and apparatus are prepared independently in initial states; it also includes situations in which we prepare a system in an initial state and then put it into interaction with an environment, such as a heat bath, that has been prepared independently.

4 The Debate Concerning Not Completely Positive Dynamical Maps

The early pioneering work of Sudarshan et al. (1961), and Jordan and Sudarshan (1961), did not assume complete positivity, but instead characterised the most general dynamical framework for quantum systems in terms of linear maps of density matrices. After the important work of, for instance, Choi (1972) and Kraus (1983), however, it became increasingly generally accepted that complete positivity should be imposed as an additional

requirement. Yet despite the reasonableness of the arguments for complete positivity, the imposition of this additional requirement was not universally accepted. Indeed, the issue of whether the more general or the more restricted framework should be employed remains controversial among physicists. At times, the debate has been quite passionate (e.g., Simmons, Jr. and Park, 1981; Raggio and Primas, 1982; Simmons, Jr. and Park, 1982).

The issues involved in the debate were substantially clarified by an exchange between Pechukas and Alicki which appeared in a series of papers between 1994 and 1995. Pechukas and Alicki analysed the dynamical map, Λ, for a system into three separate components: an 'assignment map', a unitary on the combined state space, and a trace over the environment:

$$\rho_S \rightarrow \Lambda\rho_S = \mathrm{tr}_R(U\Phi\rho_S U^\dagger), \tag{17}$$

with S, R representing the system of interest and the environment (the 'reservoir') respectively, and the assignment map, Φ, given by

$$\rho_S \rightarrow \Phi\rho_S = \rho_{SR}. \tag{18}$$

Since the unitary and the partial trace map are both CP, whether or not Λ itself is CP is solely determined by the properties of Φ, the assignment map. Φ represents an assignment of 'initial conditions' to the combined system: it assigns a *single* state, ρ_{SR}, to each state ρ_S. My use of inverted commas here reflects the fact that such a unique assignment cannot be made in general, since in general the state of the reservoir will be unknown. It will make sense to use such a map in some cases, however; for instance if there is a class Γ of possible initial states $S + R$ that is such that, within this class, ρ_S uniquely determines ρ_{SR}. Or it might be that, even though there are distinct possible initial states in Γ that yield the same reduced state ρ_S,

the evolution of ρ_S is (at least approximately) insensitive to which of these initial states is the actual initial conditions.

When Φ is linear:

$$\Phi(\lambda\rho_1 + (1-\lambda)\rho_2) = \lambda\Phi(\rho_1) + (1-\lambda)\Phi(\rho_2), \tag{19}$$

consistent:

$$\mathrm{tr}_R(\Phi\rho_S) = \rho_S, \tag{20}$$

and of product form, one can show that Φ is of necessity CP as well. Pechukas (1994) inquired into what follows from the assumption that Φ is linear, consistent, and positive. Pechukas showed that if Φ is defined everywhere on the state space, and is linear, consistent, and positive, *it must be a product map*: $\rho_S \xrightarrow{\Phi} \rho_{SR} = \rho_S \otimes \rho_R$, with ρ_R a fixed density operator on the state space of the reservoir (i.e., all ρ_S's are assigned the same ρ_R). This is undesirable as there are situations in which we would like to describe the open dynamics of systems that do not begin in a product state with their environment. For instance, consider a multi-partite entangled state of some number of qubits representing the initial conditions of a quantum computer, with one of the qubits representing a 'register' and playing the role of S, and the rest playing the role of the reservoir R. If we are restricted to maps that are CP on the system's entire state space then it seems we cannot describe the evolution of such a system.

Pechukas went on to show that when one allows correlated initial conditions, Λ, interpreted as a dynamical map defined on the entire state space of S, may be NCP. In order to avoid the ensuing negative probabilities, one can define a 'compatibility domain' for this NCP map; i.e., one stipulates that Λ is defined only for the subset of states of S for which $\Lambda\rho_S \geq 0$ (or

equivalently, $\Phi\rho_S \geq 0$). He writes:

> The operator Λ is defined, via reduction from unitary $S + R$ dynamics, only on a subset of all possible ρ_S's. Λ may be extended—trivially, by linearity—to the set of *all* ρ_S, but the motions $\rho_S \rightarrow \Lambda\rho_S$ so defined may not be physically realizable ... Forget complete positivity; Λ, extended to all ρ_S, may not even be positive (1994).

In his response to Pechukas, Alicki (1995) conceded that the only initial conditions appropriate to an assignment map satisfying all three "natural" requirements—of linearity, consistency, and complete positivity—are product initial conditions. However, he rejected Pechukas's suggestion that in order to describe the evolution of systems coupled to their environments one must forego the requirement that Λ be CP on S's entire state space. Alicki calls this the "fundamental positivity condition." Regarding Pechukas's suggestion that one may use an NCP map with a restricted compatibility domain, Alicki writes:

> ... Pechukas proposed to restrict ourselves to such initial density matrices for which $\Phi\rho_S \geq 0$. Unfortunately, it is impossible to specify such a domain of positivity for a general case, and moreover there exists no physical motivation in terms of operational prescription which would lead to [an NCP assignment of initial conditions] (Alicki, 1995).

It is not clear exactly what is meant by Alicki's assertion that it is impossible to *specify* the domain of positivity of such a map in general, for does not the condition $\Phi\rho_S \geq 0$ itself constitute a specification of this domain? Most plausibly, what Alicki intends is that *determining* the compatibility domain will be exceedingly difficult for the general case. We

will return to this question in the next section, as well as to the question of the physical motivation for utilising NCP maps.

In any case, rather than abandoning the fundamental positivity condition, Alicki submits that in situations where the system and environment are initially correlated one should relax either consistency or linearity. Alicki attempts to motivate this by arguing that in certain situations the preparation process may induce an instantaneous perturbation of S. One may then define an inconsistent or nonlinear, but still completely positive, assignment map in which this perturbation is represented.

According to Pechukas (1995), however, there is an important sense in which one should not give up the consistency condition. Consider an inconsistent linear assignment map that takes the state space of S to a convex subset of the state space of $S + R$. Via the partial trace it maps back to the state space of S, but since the map is not necessarily consistent, the traced out state, ρ'_S, will not in general be the same as ρ_S; i.e.,

$$\rho_S \xrightarrow{\Phi} \Phi\rho_S \xrightarrow{\text{tr}_R} \rho'_S \neq \rho_S. \tag{21}$$

Now each assignment of initial conditions, $\Phi\rho_S$, will generate a trajectory in the system's state space which we can regard as a sequence of CP transformations of the form:

$$\rho_S(t) = \text{tr}_R(U_t \Phi\rho_S U_t^\dagger). \tag{22}$$

At $t = 0$, however, the trajectory begins from ρ'_S, not ρ_S. ρ_S, in fact, is a fixed point that lies *off* the trajectory. This may not be completely obvious, prima facie, for is it not the case, the sceptical reader might object, that we can describe the system as evolving from ρ_S to ρ_{SR} via the assignment map and then via the unitary transformation to its final state? While this much

may be true, it is important to remember that Φ is supposed to represent an assignment of *initial conditions* to S. On this picture the evolution through time of $\Phi\rho_S$ is a proxy for the evolution of ρ_S. When Φ is consistent, $\text{tr}_R(U\Phi\rho_S U^\dagger) = \text{tr}_R(U\rho_{SR}U^\dagger)$ and there is no issue; however when Φ is inconsistent, $\text{tr}_R(U\Phi\rho_S U^\dagger) \neq \text{tr}_R(U\rho_{SR}U^\dagger)$, and we can no longer claim to be describing the evolution of ρ_S through time but only the evolution of the distinct state $\text{tr}(\Phi\rho_S) = \rho'_S$. And while the evolution described by the dynamical map $\rho'_S(0) \xrightarrow{\Lambda} \rho'_S(t)$ is completely positive, it has *not* been shown that the transformation $\rho_S(0) \xrightarrow{\Lambda} \rho_S(t)$ must always be so.

What of Alicki's suggestion to drop the linearity condition on the assignment map? It is unclear that this can be successfully physically motivated, for it is prima facie unclear just what it would mean to accept nonlinearity as a feature of reduced dynamics. Bluntly put, quantum mechanics is linear in its standard formulation: the Schrödinger evolution of the quantum-mechanical wave-function is linear evolution. Commenting on the debate, Rodríguez-Rosario et al. (2010) write: "giving up linearity is not desirable: it would disrupt quantum theory in a way that is not experimentally supported."

5 Linearity, Consistency, and Complete Positivity

We saw in the last section that there are good reasons to be sceptical with respect to the legitimacy of violating any of the three natural conditions on assignment maps. We will now argue that there are nevertheless, in many situations, good, physically motivated, reasons to violate these conditions.

Let us begin with the CP requirement. *Pace* Alicki, one finds a clear physical motivation for violating complete positivity if one notes, as Shaji and Sudarshan (2005) do, that if the system S is initially entangled with R, then not all initial states of S are allowed—for instance,

$\rho_S = \mathrm{tr}_R \rho_{SR}$ cannot be a pure state, since the marginal of an entangled state is always a mixed state. Such states will be mapped to negative matrices by a linear, consistent, NCP map. On the other hand the map will be positive for all of the valid states of S; this is the so-called compatibility domain of of the map: the subset of states of S that are compatible with Λ.

In light of this we believe it unfortunate that such maps have come to be referred to as NCP maps, for strictly speaking it is not the map Λ but its linear extension to the entire state space of S that is NCP. Λ is indeed CP *within its compatibility domain*. In fact this misuse of terminology is in our view at least partly responsible for the sometimes acrid tone of the debate. From the fact that the linear extension of a partially defined CP map is NCP, it does not follow that "reduced dynamics need not be completely positive."[2] Alicki and others are right to object to this latter proposition, for given the arguments for complete positivity it is right to demand of a dynamical map that it be CP on the domain within which it is defined. On the other hand it is *not* appropriate to insist with Alicki that a dynamical map must be CP on the entire state space of the system of interest—come what may—for negative probabilities will only result from states that cannot be the initial state of the system. Thus we believe that 'NCP maps'—or more appropriately: *Partial-CP* maps with NCP linear extensions—can and should be allowed within a quantum dynamical framework.

What of Alicki's charge that the compatibility domain is impossible to "specify" in general? In fact, the determination of the compatibility domain is a well-posed problem (cf. Jordan et al., 2004); however, as Alicki alludes to, there may be situations in which actually determining the compatibility domain will be computationally exceedingly difficult. But in other cases[3]—when computing the compatibility domain *is* feasible—we see no reason why

[2]This is the title of Pechukas's 1994 article.

[3]For examples, see Jordan et al. (2004); Shaji and Sudarshan (2005).

one should bar the researcher from using a Partial-CP map whose linear extension is NCP if it is useful for her to do so. Indeed, given the clear physical motivation for it, this seems like the most sensible thing to do in these situations.

There may, on the other hand, be other situations where proceeding in this way will be inappropriate. For instance, consider a correlated bipartite system $S + R$ with the following possible initial states:

$$x_+ \otimes \psi_+, \quad x_- \otimes \psi_-, \quad z_+ \otimes \phi_+, \quad z_- \otimes \phi_-. \tag{23}$$

The domain of definition of Φ consists of the four states $\{x_+, x_-, z_+, z_-\}$. Suppose we want to extend Φ so that it is defined on all mixtures of these states, and is linear. The totally mixed state of S can be written as an equally weighted mixture of x_+ and x_-, and also as an equally weighted mixture of z_+ and z_-.

$$\frac{1}{2}I = \frac{1}{2}x_+ + \frac{1}{2}x_- = \frac{1}{2}z_+ + \frac{1}{2}z_-. \tag{24}$$

If Φ is defined on this state, and is required to be a linear function, we must have

$$\begin{aligned} \Phi(\frac{1}{2}I) &= \frac{1}{2}\Phi(x_+) + \frac{1}{2}\Phi(x_-) \\ &= \frac{1}{2}x_+ \otimes \psi_+ + \frac{1}{2}x_- \otimes \psi_-, \end{aligned} \tag{25}$$

$$\begin{aligned} \Phi(\frac{1}{2}I) &= \frac{1}{2}\Phi(z_+) + \frac{1}{2}\Phi(z_-) \\ &= \frac{1}{2}z_+ \otimes \phi_+ + \frac{1}{2}z_- \otimes \phi_-, \end{aligned} \tag{26}$$

from which it follows that

$$\frac{1}{2}x_+ \otimes \psi_+ + \frac{1}{2}x_- \otimes \psi_- = \frac{1}{2}z_+ \otimes \phi_+ + \frac{1}{2}z_- \otimes \phi_-, \tag{27}$$

which in turn entails that

$$\psi_+ = \psi_- = \phi_+ = \phi_-, \tag{28}$$

so Φ cannot be extended to a linear map on the entire state space of S unless it is a product map.

It would be misleading to say that assignment maps such as these violate linearity, for much the same reason as it would be misleading to say that Partial-CP maps with NCP linear extensions violate complete positivity. It is not that these maps are defined on a convex domain, and are nonlinear on that domain; rather, there are mixtures of elements of the domain on which the function is undefined. But since we cannot be said to have violated linearity, then *pace* Rodríguez-Rosario et al., in such situations we see no reason to bar the researcher from utilising these 'nonlinear' maps, for properly understood, they are partial-linear maps with nonlinear extensions.

Pace Pechukas, there may even be situations in which it is appropriate to use an inconsistent assignment map. Unlike the previous cases, in this case the assignment map will be defined on the system's entire state space. This will have the disadvantage, of course, that our description of the subsequent evolution will not be a description of the true evolution of the system, but in many situations one can imagine that the description will be "close enough," i.e., that

$$\mathrm{tr}_R(U_t \rho_{SR} U_t^\dagger) \approx \mathrm{tr}_R(U_t \rho'_{SR} U_t^\dagger). \tag{29}$$

6 Conclusion

Bohr warned us long ago against extending our concepts, however fundamental, beyond their domain of applicability. The case we have just looked at is an illustration of this important point. The debate over the properties one should ascribe to the extension of a partially-defined description is a debate over the properties one should ascribe to a phantom.

Whether or not we must use a map whose extension is nonlinear, or a map whose linear extension is NCP, or an inconsistent map, is not a decision that can be made a priori or that can be shown to follow from fundamental physical principles. The decision will depend on the particular situation and on the particular state preparation we are dealing with.

References

Alicki, Robert. "Comment on 'Reduced Dynamics Need Not Be Completely Positive'." *Physical Review Letters* 75 (1995): 3020.

Choi, Man-Duen. "Positive Linear Maps on C*-Algebras." *Canadian Journal of Mathematics* 24 (1972): 520–529.

Devi, A. R. Usha, A. K. Rajagopal, and Sudha. "Quantumness of Correlations and Entanglement." *International Journal of Quantum Information* 9 (2011): 1757–1771.

Jordan, Thomas F., Anil Shaji, and E. C. G. Sudarshan. "Dynamics of Initially Entangled Open Quantum Systems." *Physical Review A* 70 (2004): 052,110.

Jordan, Thomas F., and E. C. G. Sudarshan. "Dynamical Mappings of Density Operators in Quantum Mechanics." *Journal of Mathematical Physics* 2 (1961): 772–775.

Kraus, Karl. *States, Effects, and Operators*. Berlin: Springer-Verlag, 1983.

Nielsen, Michael A., and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press, 2000.

Pechukas, Philip. "Reduced Dynamics Need Not Be Completely Positive." *Physical Review Letters* 73 (1994): 1060–1062.

———. "Pechukas Replies." *Physical Review Letters* 75 (1995): 3021.

Raggio, G.A., and H. Primas. "Remarks on "On Completely Positive Maps in Generalized Quantum Dynamics"." *Foundations of Physics* 12 (1982): 433–435.

Rodríguez-Rosario, César A., Kavan Modi, and Alán Aspuru-Guzik. "Linear Assignment Maps for Correlated System-Environment States." *Physical Review A* 81 (2010): 012,313.

Shaji, Anal, and E. C. G. Sudarshan. "Who's Afraid of Not Completely Positive Maps?" *Physics Letters A* 341 (2005): 48–54.

Simmons, Ralph F., Jr., and James L. Park. "On Completely Positive Maps in Generalized Quantum Dynamics." *Foundations of Physics* 11 (1981): 47–55.

———. "Another Look at Complete Positivity in Generalized Quantum Dynamics: Reply to Raggio and Primas." *Foundations of Physics* 12 (1982): 437–439.

Sudarshan, E. C. G., P. M. Mathews, and Jayaseetha Rau. "Stochastic Dynamics of Quantum-Mechanical Systems." *Physical Review* 121 (1961): 920–924.

Narratives & Mechanisms

Abstract

Historical scientists are frequently concerned with narrative explanations targeting single cases. I show that two distinct explanatory strategies are employed in narratives, *simple* and *complex*. A simple narrative has minimal causal detail and is embedded in a general regularity, whereas a complex narrative is more detailed and not embedded. This distinction's importance is illustrated in reference to mechanistic explanation. I consider 'liberal' accounts of mechanistic explanation, which expand the traditional picture to accommodate less mechanistic sciences. Simple narratives warrant a mechanistic treatment, while some complex narratives do not.

Introduction

Scientists examining the past are taken to be primarily concerned with *narrative* explanations which account for single events[1]. A meteor exterminated the dinosaurs; New Zealand's lake Taupo was formed by an enormous volcanic eruption; the introduction of small-pox killed millions in the Americas. Of course, historical scientists are not narrowly concerned with narrative explanation. As Kosso (2001) and Jeffares (2008) discuss, they sometime target middle-range theories which connect contemporary phenomena to past events (see also Turner 2009). Moreover, much historical enquiry targets patterns and regularities in deep time. Paleobiological work covering the nature of mass extinction events (Raup 1991) the nature of speciation (Eldredge & Gould 1972), the role of selection and adaptationist explanations in macro-level patterns (Gould et al 1977, Huss 2009), are all concerned with regularities in life's shape, not the explanation of a simple event. However, at least much of the time their explanatory interests are geared towards the particular rather than the general. This paper shows that historical explanation, understood as narrative, is disunified: at least two distinct explanatory strategies are employed. *Simple* narratives explain particular cases as instances of regularities – the explanandum is subsumed by a general model. *Complex* narratives do not account for explananda in terms of regularities or models.

I argue that simple narratives have more in common with the population-level explanations furnished by economists and ecologists than complex narratives. This is demonstrated by comparing narrative explanations with mechanistic models. Both population-level and simple narratives are amenable to

[1] For example, Kitcher 1993, Cleland 2011, Hempel 1965 and Hull 1975 appear to agree that historical enquiry is primarily narrative

mechanistic gloss. However, in complex cases scientists are not typically mechanists. Faced with a complex world, they employ characteristically non-mechanistic explanations.

The paper is in three parts. In the first, two case-studies illustrate the distinction between simple and complex narratives. Part two discusses mechanistic explanation, sketching the view and introducing liberalism – the view that most or all scientific explanation is mechanistic. The third part examines narrative explanation in light of mechanistic explanation, arguing that simple narratives are characteristically mechanistic, while some complex narratives are not.

1. Narrative Explanations

Narrative explanations account for particular events[2] via causal sequences concluding with the explanandum. The causal sequence makes the explanandum likely. Narrative explanations are taken to be distinctively historiographical (at least by Hempel and Hull) due to their 'story-like' structure and lack of appeal to laws. The treaties at the close of the First World War led inevitably to the Second; the extraterrestrial impact which caused the Chicxulub crater was sufficient to exterminate the dinosaurs; and so on. There is more than one way to account for an event, however. Some causal sequences stand alone: even if only one extinction event was caused by an impact, we can be convinced of the impact's causal sufficiency. Or we might explain an event as an instance of a general model: perhaps all wars have common causes, and the Second World War can be explained in terms of those commonalities.

I will be agnostic as to whether all narratives in fact reference regularities, and whether this is problematic. Hempel's primary concern about historiographic explanation is the lack of nomological appeals and I (in part) share the suspicion that particular events can be satisfactorily explained without recourse to regularities (c.f Tucker 1998) but my claim of the disjunctive nature of narratives holds regardless of this.

Hopefully it is clear that narrative explanations are surely not restricted to historiographical inquiry – there is nothing stopping a chemist explaining a single event in terms of some causal sequence (perhaps even without explicit mention of laws) - and therefore the claims I make about narrative explanation will most likely not be restricted to the geological and paleontological cases I focus on. Whether the distinctions and lessons I draw are extendable to other sciences I leave for future work: given that

[2] I will speak in terms of past events, but historical enquiry also covers historical processes, entities and states of affairs. The claims made about events carry over to those other types of targets.

narrative explanation is paradigmatically the business of historical inquiry, it is the obvious place to center philosophical investigations

And so narrative explanations (1) account for some particular explanandum in terms of some causal sequence; (2) may or may not appeal explicitly to laws or generalizations; (3) are paradigmatically, but not exclusively, historical. I argue that there are two explanatory strategies which historical scientists employ in providing narratives.

1.1 Snowball Earth

There were glaciers in the tropics at least twice during the Neoproterozoic (roughly 1000 – 542 million years ago). Towards the end of the period there was synchronous, ubiquitous glaciation: the entire earth covered in permafrost cut through by rivers of ice. This presents a series of geological and palaeoclimatological challenges. What could have caused this scenario? How did it thaw? Why are such events rare? The most popular explanation of these glacial events is Joseph Kirschvink's Snowball Earth Theory (Schopf & Klein 1992, Hoffman & Schrag 2002).

The late Neoproterozoic was a time of continental dispersal: the supercontinent Rodinia broke up and the megacontinent Gondwana began to form. During glacial periods most continents clustered at the middle and lower latitudes. Kirshvink proposed that this clustering was responsible for the global freeze.

Both land and ice-caps have high albedo – they reflect more of the sun's energy than water. Tropical landmasses have high albedo because more sunlight reaches the equator. Their warm, moist climate also increases silicate weathering (the absorption of $C0_2$). Land clustering around the tropics, then, increases albedo and decreases greenhouse gases. This would lower the earth's temperature – particularly at the poles where the growth of ice sheets would lead to a freezing feedback loop:

> If more than about half of the Earth's surface area were to become ice covered, the albedo feedback would be unstoppable… surface temperatures would plummet, and pack ice would quickly envelope the tropical oceans (Hoffman & Schrag pp 135)

The explanation can be presented in a simple flowchart:

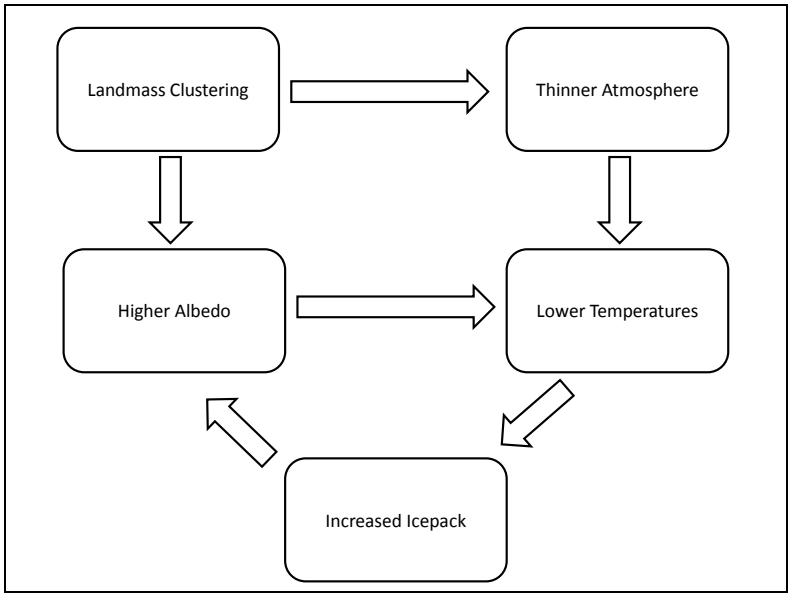


Figure 1: Snowball Earth

Landmass clustering in the tropics lowers temperature by increasing albedo and thinning the atmosphere. Lower temperatures increase icepack cover, creating a feedback loop between lowering temperatures, larger icecaps, and higher albedo. Earth freezes over. As we shall see, Snowball Earth is a paradigm 'simple' narrative: an event is explained by a general model with reference to minimal causal factors.

1.2 Sauropod Gigantism

Despite public perception, most dinosaurs fit comfortably in the familiar mammalian size-range. The sauropods were different: not merely big, but puzzlingly so. Some were the largest land animals to have ever lived: *Sauroposeidon* and *Argentinosaurus* are estimated to have weighed between 50 and 70 tons, rivaling baleen whales in length. By contrast, the largest known terrestrial mammal was *Paraceratherium*, thought to be 12 meters long and weighing 20 tons at most. How did sauropods manage such sizes? Why was it unique? How was gigantism physiologically and evolutionarily possible?

As Sander, Christian et al (2011) review, sauropod gigantism was the result of myriad causes (see also Klein et al 2011). Sauropods were the right lineage, in the right place, at the right time. They had specific primitive characteristics which removed size limitations. Early sauropods were oviparous – egg-laying

allows for fast population recovery, mitigating the small population size engendered by gigantism. They did not masticate, increasing food intake. They had a distinctive small-head-and-long-neck morphological structure, which maximizes grazing range while minimizing movement.

These primitive characteristics were supplemented by new adaptations. Gigantism itself protected against the increasingly sophisticated predators of the Jurassic, and accommodated the enormous digestive system mitigating the lack of mastication and gastric mill. Their basal metabolic rate increased to accommodate the speedy growth required. Sauropods evolved a distinctive pneumatized skeleton, a signal of a bird-like respiratory system, which increases the efficiency of oxygen dispersal and accommodates the growth rate required to reach gigantic size.

The road to gigantism was open to sauropods due to their distinctive primitive characteristics. The road was followed due to the evolution of particular adaptations in response to particular evolutionary pressures. The explanation of sauropod gigantism is a complex narrative: there is no appeal to a general model in explanation, but rather a unique, detailed causal sequence is employed.

1.3 Simple & Complex Narratives

In explaining snowball earth and sauropod gigantism historical scientists follow two distinct explanatory strategies. Both are narrative explanations: their explananda are individual cases, accounted for via particular causal sequences. However, snowball earth is explained as an extreme case of a general model. Sauropod gigantism is not. Moreover, the Snowball Earth contains less causal detail than sauropod gigantism. The geological case is *simple,* while the paleobiological case is *complex*[3]. Two features, an explanation's *detail* and *embeddedness*, are characteristic of simple and complex narratives. It is worth reiterating that these distinctions may well illuminate sciences not typically considered historical, or dealing with narratives. It is beyond this paper's scope to discuss such cases, but I take it that if simple and complex explanations of particular events occur in ahistorical sciences, this only strengthens the importance of the distinction.

Detail

A striking difference between the two explanations is the level of *detail* required. Detail is a measure of the specificity, complexity and diffusion of the explanans required for explanatory adequacy. Snowball

[3] The distinction between complex and simple is similar in spirit to 'actual sequence' and 'robust process' explanations (Sterelny 1996, Jackson & Pettit 1992), although is not cashed out in overtly modal terms.

earth is low- detail: few factors and a single difference-maker are required. General facts about global albedo, temperature, atmosphere and icepack work in tandem with particular facts about landmass clustering to produce the explanandum. Sauropod gigantism, by contrast, requires a more detailed explanation. Adequacy requires many explanans, quite disparate in nature. Important explanatory details are spread through time: from deeply primitive characteristics such as oviparity, to highly derived ones like pneumatization. Explanans are also spread across grain: oviparity is important because it mitigates evolutionary, population-level concerns while pneumatization solves individual-level, physiological concerns.

Detail, then, tracks the complexity required for explanatory adequacy, and its nature depends in part on the explanandum. In the snowball earth case, the world cooperates in granting sufficiency to low detail explanations while for sauropod gigantism the distended, messy nature of the explanandum demands a more detailed, messy explanation.

Embeddedness

A narrative explanation is *embedded* when the explanandum is accounted for as a token of a type of process; an instance of a regularity. The relative simplicity of the snowball earth explanation allows it to be represented by a single climatological model. The hypothesis is an extreme case of run of the mill dynamics between ice cover, geography, climate and atmosphere. In explaining why the earth froze, I tell you about those general dynamics and how the scenario would arise given particular states of affairs. Sauropod gigantism, by contrast, is an exquisite corpse: birds provide a model for respiratory systems; giraffes, swans and structural morphology tell us something about possible sauropod stances; elephants and large lizards about possible metabolism. There is no single unifying regularity which can be appealed to. In explaining gigantism, I refer to particular facts about the sauropod lineage and the environment in which it evolved.

I have mentioned that some philosophers take narrative explanations as problematic insofar as they do not appeal to regularities, and that I will not take a stance on this here. With embeddedness on the table, I can clarify this. Clearly embedded explanations appeal to regularities: the interesting question is whether non-embedded explanations do, or must. I am inclined to see non-embedded explanations as leaning on a patchwork of regularities. For instance, models of structural morphology, population genetics and metabolism are all appealed to in explanations of sauropod gigantism. However, it is open for others to argue that such appeals are not always required.

Embeddedness, then, tells us whether an explanandum is accounted for as an instance of a general model, or as an individual event.

Simple & Complex

Call an explanation which is high in detail, and not embedded, a *complex narrative*. Call an explanation with is low in detail and embedded a *simple narrative*. Complex and simple narratives are two distinct explanatory strategies employed by historical scientists.

To drive the distinction home, compare the explanation of gigantism in sauropods to cases of island gigantism. The six-foot, tree climbing, predatory Fossa of Madagascar, for instance, evolved from much smaller mongoose-like ancestors. Because islands are isolated and tend to lack diversity, diminutive lineages are likely to form founder populations and radiate into unusual niches. This can lead to island gigantism: a lack of predation, and selection pressure to fill empty niches, drives size increase. The Fossa are gigantic because the isolation of Madagascar set up the preconditions for island gigantism. Fossa are amenable to a simple explanation: embedded in general explanations of island biogeography and requiring minimal detail. To explain fossa gigantism, I need only explain the general model of island gigantism, and then show how fossa met the model's conditions for evolving large size. Sauropod gigantism, by contrast, begs a complex explanation: more detail is required and there is no general regularity to subsume the explanandum.

Detail and embeddedness come apart in principle, but in practice tend to be coupled. Embedded explanations tend to be low in detail as explanatory sufficiency is determined by the strictures of the model. To get the Snowball Earth explanation, I show that the antecedent conditions of the model were met – and this only requires reference to causal factors from that model. This allows many causal details to be ignored, making for a low-detail explanation. Non-embedded explanations tend to require more detail as they cannot rely on general regularities to discount causal factors. In the Sauropod case, we require separate convincing of each step in the explanation. There may be cases of embedded, high detail explanations as well as low detail, unembedded explanations, but these are rare.

A *simple* narrative explanation, then, does not require a detailed treatment as the explanandum is represented in a general model. A *complex* narrative requires specific details unique to the case at hand and is not subsumed under a particular model.

This distinction is important. First, it explains two divergent approaches to understanding the explanatory unity of narratives. In Hull's treatments (1975, 1989) narrative explanations owe their unity in part to the integrity of the *historical entity* they target. "The role of the central subject is to form the main strand around which the historical narrative is woven (255)." According to Hull, accounts of sauropod gigantism and snowball earth are explanatory in virtue of picking out central subjects (spatio-temporally distended objects), and providing a coherent narrative about that subject.

By contrast, Glennan (2010) argues that narrative explanations operate through *ephemeral mechanisms.* By his lights, historical scientists explain states of affairs by showing that the preconditions for a general mechanism are in place. Such mechanisms are unusual due to their contingent (hence 'ephemeral') nature, but still deliver robust results *given* that arrangement. The characteristics of early sauropods, or the continental arrangement of the Neoproterozoic, are highly contingent states of affairs. But *given* those states of affairs, we get general results: gigantism and a general freeze (see Gallie 1959 for a similar view).

For Hull then, part of a narrative's explanatory unity is due to their central subject. For Glennan, unity is owed to regularities. This disagreement is resolved when we see that narrative explanations take two different forms. In simple cases, historical scientists appeal to general models which subsume the target case as Glennan envisions. In complex cases, explanatory force might be supplied by historical entities as Hull sees it.

Second, the distinction shows that historical scientists are not unified in their approach to explanation. They pursue two distinct strategies which require separate philosophical treatments. I illustrate this in reference to mechanistic explanation. It turns out that simple narratives can receive a mechanistic gloss, while some complex narratives are not mechanistic. Showing this is the task of the second half of the paper.

2. Mechanistic explanation

Mechanistic explanation has proven an illuminating account of actual scientific practice (see Bechtel & Richardson 1993, Glennan 2002, Craver 2007, Woodward 2002, Machamer, Darden et al 2000) In this section I sketch the account, then discuss how it may be extended to cover population-level explanation. Discussing narrative explanation in the context of mechanistic explanation will show that 1) simple narrative explanations are unified with population-level explanations (but not with complex narrative explanations) and 2) not all scientific explanations are mechanistic (as some complex narratives are not).

There are reasons to compare mechanistic and narrative explanation. First, there is a tension between historical explanation and models of explanations referring to hierarchical structure, such as reduction. Traditional models of reduction require explanation to refer to general laws which are realized at 'more fundamental' levels of description than the explanandum. Such laws are not overtly appealed to in historical explanation. Mechanistic explanation is intended to replace reductive models (Craver 2005, Bechtel & Abrahamsen 2005), retaining their advantages but avoiding imperialistic and nomological pitfalls. If mechanistic explanation is such a replacement, we might wonder whether the tension between historical and structural explanation is retained.

Second, as 2.2 covers, there is interest in the limits of mechanistic explanation. Is mechanistic explanation a general account of scientific explanation, or is it one of many explanatory strategies scientists might follow?

As we shall see, the tension between historical and structural explanation is retained in some complex narratives. In such cases scientists do not attempt mechanistic explanations because the unembedded, high-detail nature of the explanation undermines the utility of a mechanistic approach. And for the same reason mechanistic explanation has limited scope: scientists are not just in the mechanism business, sometimes they are in the complex narrative business.

Third, understanding the nature of historical explanation is a worthy philosophical task and its relationship to mechanistic accounts is illuminating. I have already shown that narrative explanation is disjunctive between simple and complex strategies. As we shall see, simple narratives are unified with population-level explanations via their common 'mechanistic' nature, while complex narratives are the odd ones out.

2.1 A sketch

In this section I aim to provide a minimal set of conditions required for any explanation to be presented mechanistically. In explaining a mechanism I must identify the phenomenon I am concerned with, break it into components, and explain the phenomenon's behavior in terms of the causal and organizational properties of the components. For the purposes of this paper, I will take an explanation to be mechanistic if it meets the following criteria:

1) *Localization*: the phenomenon is a discrete system with discrete components

2) *Constitution*: systems are constitutively explained in terms of components

3) *Nested Causation*: behaviors of systems are explained in terms of the causal and relational properties of components

This sketch is certainly not exhaustive of all that is important and distinctive about mechanistic explanation. However, it is a minimal set of conditions which I hope mechanists of all stripes would agree with and are all I need for present purposes. It is clear that more needs to be said about localization: what is meant by 'discrete', and how does it restrict the scope of mechanistic explanations? I will put this question aside until 2.2.

Many sciences are characteristically mechanistic. Cytologists understand cells as discrete parcels composed of a cellular anatomy which determines behavior. Neuroscientists identify neural networks as systems fulfilling particular functions governed by activation patterns within them. Molecular geneticists identify genes with particular DNA sequences which code for proteins given the right inputs and organization. Chemists explain phase-transitions as the result of the interaction between kinetic energy and chemical bonds in a system. All follow mechanistic explanation's distinctive pattern.

However, some scientific endeavors look different. Ecologists, economists and evolutionary biologists use abstract models to explain the behavior of populations. Paleontologists, geologists and archaeologists construct narrative explanations of events in the deep past. Using abstract models to explain population-level phenomena and using causal sequences to explain past states of affairs appear very different from the explanations mechanists examine. In the next section, we see whether mechanistic explanation can account for these as well.

2.2 'Liberalism' about mechanistic explanation

Consider two views on the scope of mechanistic explanation. By a *conservative* view the model has thin scope - it is true of some, but not all, scientific explanations. A *liberal* view takes the model to have wide scope – most, perhaps all, scientific explanations are mechanistic. Liberalism involves showing that various explanatory schema are subsumed by mechanistic accounts, and this may involve tweaking the conditions sketched above.

Let's start with two examples. Bechtel (2011) argues that mechanistic explanation must include dynamic causal streams to capture biological phenomena which display non-linear behavior, such as cellular self-repair. It is not obvious that mechanists ever intended their models to be rigidly linear, and moreover expanding the account to include dynamic mechanisms doesn't seem to conflict with anything essential

to the sketch above. By contrast, Rusanen and Lappi (2007) argue that some cognitive phenomena are beyond the scope of mechanistic models as they require top-down explanation. This clashes with constitution: instead of the phenomenon being explained in terms of its parts, the parts are explained via the phenomenon. If they are right, mechanists have a choice between the conservative move of taking some cognitive explanations non-mechanistically, or the liberal move of altering the requirement of constitution. Some cases, then, are more or less challenging to the model.

A liberal move pertinent to comparing narrative and mechanistic explanation is discussed by Matthewson & Calcott (2011). They argue that explanations of population-level phenomena, such as market cycles and predator/prey dynamics, can be understood mechanistically. I argue that simple narrative explanations can be understood in the same way.

Matthewson & Calcott distinguish between *mechanisms* and *mechanistic models*. A mechanism is a concrete object with localizable, discrete components. A mechanistic model takes the *structure* of mechanistic explanation and applies it to non-mechanisms. It is not clear whether economies, ecologies or cities are mechanisms, but we may successfully explain them *as if they were*. In explaining their target, modelers entertain the fiction (Godfrey-Smith 2009) or the idealization (Weisberg 2007) that it is a mechanism, enabling them to employ mechanistic explanation.

Take an evolutionary explanation of a shift in the proportion of some trait, *t*, in a population across two subsequent generations, G_1 and G_2. In G_1 *t* is less common than it is in G_2. To explain this change, a biologist might refer to a model which considers the population in terms of various traits with various fitness-values. The makeup of the population at one generation is determined by the fitness values of the traits present in the generation before. Because of *t*'s fitness value, it outperformed some other traits in reproducing between G_1 and G_2 and was thus more common in the later generation. Whether this is a mechanistic explanation depends upon its interaction with the conditions I outlined above.

The explanation is mechanistic, with a tweak. First, it involves decomposition: the population is understood as comprising either individuals or traits with fitness-values. Second, it involves nested causation[4]: the change between the two generations is explained as the result of the interacting fitness

[4] This example is meant to be illustrative, and skates over some difficult issues in biology. Some philosophers (Walsh, Lewens & Arieu 2002; Walsh 2010) deny that fitness is truly causal, insisting that only the particular life-events of individuals in the population are the proper locus of causal power – and thus calling this 'nested causation' is a mistake. Fair enough, but I think this perspective is in fact amenable to the story I am telling. First,

values of the components. However, the phenomenon does not appear to be a discrete system. Few real-world biological populations have discrete, non-overlapping generations and even fewer have populations as discrete as the model represents. And yet the system is treated *as if* it were a discrete, localizable system. Matthewson & Calcott can retain the first tenet by allowing for idealized, or metaphorical localization. Something like:

> *Localization**: the phenomena either is a discrete system, or may be *treated like* a discrete system

Until now I have avoided explicit discussion of what is meant by 'discreteness', but it is time to draw this out. A discrete system is not necessarily such in virtue of spatio-temporal location, but rather the *causal integration* of its parts. It has discrete components insofar as they are *modular*: they perform particular, identifiable and perhaps extractable functions in the context of that system (this account is meant to be broadly aligned with Wimsatt's (2007)). A clockwork machine can be a paradigmatically discrete system. It is discrete in terms of causal integration: the behaviors of clockwork (keeping time, say) depend upon the interaction of a specific set of contained parts. Moreover, the components are modular: the various cogs and wheels can be removed from the system and play identifiable roles within it. When Matthewson & Calcott argue that population-level explanations are capturable by mechanistic models, they simply idealize from a paradigmatically discrete system, to a less clear case.

'Discreteness', as I understand it, is clearly graded; and this should make *localization** unproblematic for mechanists – most accounts of mechanistic explanation already commit to something like this. Indeed, discussion of mechanistic explanation is rife with discussion of idealization. And so a clockwork machine is quite discrete. A neural network is less so: although neuroscientists individuate networks via examining neuroanatomy and firing patterns, complex overlapping and interrelation exists between the entities in the system. The more the example diverges from an ideally discrete system, the more metaphorical in character the mechanistic explanation of it becomes. This has consequences for the process of localization applied in different cases. For more 'machine-like' cases, such as clockwork, we

---

one might claim that non-causal factors are here presented *as if* they were causal, and so 'nested causation' is, like localization, receiving a fictionalist treatment. Second, one could claim that 'fitness' in the this context is merely a term of art meant to unite whatever truly causal factors in fact lead to the births and deaths which occur within the population. Moreover, the main concern of such philosophers is whether explanations appealing to fitness should be read as *mathematical* explanations – and discussion of the relationship between mathematical and mechanistic explanations is beyond the scope of this essay.

are more able to 'read' the system from the world. The components of the mechanistic model map onto components in the world. In less 'machine-like' cases, a process of simplification, abstraction or idealization is required. The 'fitness value' of some trait, for instance, does not obviously (if at all) map onto components in the real world system. They rather pick out explanatorily salient features of the target. Representing population-level phenomena as discrete systems requires that we ignore certain causal factors. This is not, of course, an original claim – indeed I think it is necessary for understanding mechanistic explanations, but it is worth restating for as we shall see, although such idealizations occur in simple narrative explanations, they do not in many complex cases.

And so Matthewson & Calcott are able to present many of the explanations in population-level science as mechanistic insofar as they accept a 'fictionalist' turn in localization. Given that many paradigm examples of mechanistic explanation (neural networks, gene sequences) are themselves only ideally discrete this change is not too problematic. However, the process of localization changes depending on the discreteness of the system: for characteristically mechanistic phenomena the system can be 'read off' the world, for other cases a process of simplification is required. The final section brings this liberal account of mechanistic explanation together with narrative explanation.

3. Mechanistic Narratives?

Let's take stock. Narrative explanations, which explain individual events via causal sequences, take two distinct strategies:

Simple narratives, which 1) explain an event as a state of a general model, 2) contain minimal detail;

Complex narratives, which 1) explain the event via a unique causal sequence, 2) are highly detailed.

To be mechanistic, an explanation must meet three criteria: *localization*, *constitution* and *nested causation.* Via a fictionalist tweak to localization, population-level explanations can be seen as mechanistic.

This section argues that 1) simple narrative explanations are mechanistic in the same sense as population-level explanations as they are an instance of the same explanatory strategy; 2) some

complex narratives are not mechanistic. The upshot of these two points is that liberalism about mechanisms is restricted (as there are scientific explanations which are not mechanistic) and that simple narratives have more in common with non-historical explanations (such as those from economics, sociology and ecology) than complex narratives.

3.1 Simple Narratives as Mechanistic Models

Population-level explanations and simple narrative explanations are unified. The economist treats real world markets as if they were discrete mechanisms, the evolutionary biologist imagines an island eco-system as constituted by various ecological roles waiting to be filled by genealogical actors. My exemplar simple narrative, Snowball Earth, is also an exemplar mechanistic model.

An explanatory model is mechanistic when it meets the three criteria, with the fictionalist turn described in 2.2. The phenomenon must be treated as if it were a discrete system with discrete components. It must be described constitutively. Its behavior must be explained as the result of interactions between its components. Consider the explanation sketched in 1.1. Presumably the real-world interrelation between ice-cover, atmosphere and global temperatures are extremely complex. The explanation, however, is straightforward: paleoclimatologists are able to abstract from the details and present a simple model of the interactions. The highly interrelated, complex system is treated *as if* it were a simple, discrete system. Localization holds. This idealized system is constituted by various components, namely: global temperature, icepack cover, the locations of landmasses and global albedo. Constitution holds. And the system's behavior is ruled by the causal relationships between those components. As albedo increases and the atmosphere thins due to landmasses clustering around the equator, a feedback involving decreasing temperature, increasing ice cover, and increasing albedo leads to a snowball earth scenario. Nested causation holds.

Although simple narratives and population-level models are both instances of the same explanatory strategy, it does not follow that scientists concerned with discovering historical facts face identical epistemic challenges, or use the same methods, as sciences concerned with population-level facts. It may be that ecologists and economists employ the modeler's strategy to deal with the over-abundance of facts pertaining to their explananda, while historical scientists use it to gain access to the scarcity of traces available from the past. My point is about the unity of explanatory strategies.

3.2 Complex Narratives are not Mechanistic

Scientists providing complex narratives do not attempt to embed their explanations in an overarching system, but rather provide a causal sequence which reasonably leads to the state of affairs in question. Typically, scientists providing complex narratives do not describe a localized system and do not take the explanans as system-components. They are not mechanistic.

Sauropod gigantism could in principle be explained via a 'gigantism mechanism' whereby a diminutive lineage is fed into a massively complicated idealized machine, outputted as giants millions of years later. But scientists do not explain them in those terms. Rather the history of a particular lineage is explained in reference to various causal factors interacting with it. Geologists explain Snowball Earth by representing the target as a mechanistic model. It is simplified to a localized system. There is less simplification in the sauropod case: scientists do not see the lineage as a system. After all, what would such a system look like? The explanans paleobiologists appeal to are at many temporary and hierarchical grains, and it is not obvious whether such a disparate group is amenable to unified representation. Moreover, there is a difference between a unified model and a conjunction of different (perhaps incommensurable) models. Explanans are not 'components' but rather causal factors which influenced the particular pathway the lineage took.

The point is this: even if the explanation can be described in mechanistic terms, that is not the explanation's form.

Why not? The process of localization is opaque for complex narratives due to a tension between providing a simple, tractable model and meeting the high-detail requirements of the explanation. Mechanistic approaches are attractive when the world cooperates: either the explanandum is a discrete, decomposable system or it is simple and unified enough to be helpfully treated as such. In at least some complex narratives, the requirement for high detail and the unavailability of a general regularity conspire to undermine the utility of a mechanistic conception.

Historical scientists, then, are not always mechanists. Faced with a complex, messy world they sometimes respond with complex, messy explanations.

Conclusion

I have argued that historical scientists follow two distinct explanatory strategies. Simple narratives typically idealize and abstract away from their target and are amenable to a mechanistic gloss. Complex narratives are different: some do not admit of mechanistic treatments. When providing complex narratives, historical scientists are not mechanists. This points to a host of new questions. Are there situations when simple or complex approaches are more appropriate? I have suggested that the nature of explananda play an important role in applicability, but much more remains to be said. Ought we prefer simple or complex narratives? I have said nothing about the value such explanations have. I am inclined to think of the strategies as geared towards different explanatory interests and kinds of explananda, and so validity turns on context. However, the floor is still open for those who prefer one over the other. Historically philosophers have preferred the kind of unified explanations offered by simple narratives but in some cases complex narratives may be more testable. As Kim Sterelny has pointed out to me (personal communication), a detailed narrative will have more points of empirical contact with the world, and so may have more opportunities for testing.

Finally, do other sciences have similar divisions? I have presented a unified picture of some of the explanations furnished by ecologists and economists on the one hand, and paleontologists and geologists on the other. It will be interesting to see whether some population-level explanations diverge from this pattern, and whether other areas of science can be carved up along similar lines. Moreover, I have not claimed that narrative explanations are unique to historical science (although they may be paradigmatic of them), and an investigation into whether the distinction between complex and simple narratives is useful outside of that context is also in the offing.

Attending to the different strategies historical scientists employ in their explanations illuminates important philosophical issues, and helps us understand the nature of their work.

Bibliography

Bechtel, W. (2011). "Mechanism and Biological Explanation." Philosophy of Science **78**(4): 533-557.

Bechtel, W. and A. Abrahamsen (2005). "Explanation: A mechanist alternative." Studies in History and Philosophy of Biol and Biomed Sci **36**(2): 421--441.

Bechtel, W. and R. C. Richardson (1993). Discovering complexity : decomposition and localization as strategies in scientific research. Princeton, N.J., Princeton University Press.

Cleland, C. E. (2011). "Prediction and Explanation in Historical Natural Science." The British Journal for the Philosophy of Science.

Craver, C. F. (2005). "Beyond reduction: mechanisms, multifield integration and the unity of neuroscience." Stud Hist Philos Biol Biomed Sci **36**(2): 373-395.

Craver, C. F. (2007). Explaining the brain : mechanisms and the mosaic unity of neuroscience. Oxford New York Oxford University Press,, Clarendon Press ;.

Eldredge, N. and S. J. Gould (1972). Punctuated Equilibria: an alternative to phyletic gradualism. Models in paleobiology. T. J. M. Schopf. San Francisco,, Freeman**:** vi, 250 p.

Gallie, W. B. (1959). Explanations in history and the genetic sciences. Theories of history; readings from classical and contemporary sources. P. L. Gardiner. Glencoe, Ill., Free Press**:** 549 p.

Glennan, S. (2002). "Rethinking mechanistic explanation." Philosophy of Science **69**(3): S342-S353.

Glennan, S. (2010). "Ephemeral Mechanisms and Historical Explanation." Erkenntnis **72**(2): 251-266.

Godfrey-Smith, P. (2009). "Models and fictions in science." Philosophical Studies **143**(1): 101-116.

Gould, S. J., D. M. Raup, et al. (1977). "The Shape of Evolution: A Comparison of Real and Random Clades." Paleobiology **3**(1): 23-40.

Hempel, C. G. (1965). Aspects of scientific explanation, and other essays in the philosophy of science. New York,, Free Press.

Hoffman, P. F. and D. P. Schrag (2002). "The snowball Earth hypothesis: testing the limits of global change." Terra Nova **14**(3): 129-155.

Hull, D. L. (1975). "Central Subjects and Historical Narratives." History and Theory **14**(3): 253-274.

Hull, D. L. (1989). The metaphysics of evolution. Albany, State University of New York Press.

Huss, J. (2009). The Shape of Evolution: the MBL model and clade shape. The paleobiological revolution : essays on the growth of modern paleontology. D. Sepkoski and M. Ruse. Chicago, University of Chicago Press**:** 568 p.

Jackson, F. and P. Pettit (1992). "In Defense of Explanatory Ecumenism." Economics and Philosophy **8**(1): 1-21.

Jeffares, B. (2008). "Testing times: regularities in the historical sciences." Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences **39**(4): 469-475.

Kitcher, P. (1993). The advancement of science : science without legend, objectivity without illusions. New York, Oxford University Press.

Klein, N. (2011). Biology of the sauropod dinosaurs : understanding the life of giants. Bloomington, Ind., Indiana University Press.

Kosso, P. (2001). Knowing the past : philosophical issues of history and archaeology. Amherst, N.Y., Humanity Books.

Machamer, P., L. Darden, et al. (2000). "Thinking about mechanisms." Philosophy of Science **67**(1): 1-25.

Matthewson, J. and B. Calcott (2011). "Mechanistic models of population-level phenomena." Biology and Philosophy **26**(5): 737-756.

Raup, D. M. (1991). Extinction : bad genes or bad luck? New York, W.W. Norton.

Rusanen, A.-M. and O. Lappi (2007). The Limits of Mechanistic Explanation in Neurocognitive Sciences. Proceedings of the European Cognitive Science Conference 2007. D. K. a. A. P. S. Vosniadou, Lawrence Erlbaum Associates.

Sander, P. M., A. Christian, et al. (2011). "Biology of the sauropod dinosaurs: the evolution of gigantism." Biological Reviews **86**(1): 117-155.

Schopf, J. W. and C. Klein (1992). The Proterozoic biosphere : a multidisciplinary study. Cambridge ; New York, Cambridge University Press.

Sterelny, K. (1996). "Explanatory pluralism in evolutionary biology." Biology & Philosophy **11**(2): 193-214.

Tucker, A. (1998). "Unique Events: The Underdetermination of Explanation." Erkenntnis **48**(1): 61-83.

Turner, D. (2009). Beyond Detective Work: Empirical Testing in Paleontology. The paleobiological revolution : essays on the growth of modern paleontology. D. Sepkoski and M. Ruse. Chicago, University of Chicago Press**:** 568 p.

Walsh, D. M. (2010). "Not a sure thing: Fitness, probability, and causation." Philosophy of Science **77**(2): 147-171.

Walsh, D. M., T. Lewens, et al. (2002). "The trials of life: Natural selection and random drift." Philosophy of Science **69**(3): 452-473.

Weisberg, M. (2007). "Three kinds of idealization." Journal of Philosophy **104**(12): 639-659.

Wimsatt, W. C. (2007). Re-engineering philosophy for limited beings : piecewise approximations to reality. Cambridge, Mass., Harvard University Press.

Woodward, J. (2002). "What is a mechanism? A counterfactual account." Proceedings of the Philosophy of Science Association **2002**(3): S366-S377.

Does Structural Realism Provide the Best Explanation of the Predictive Success of Science?

Gerald Doppelt
Department of Philosophy
U. of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093

jdoppelt@ucsd.edu

Abstract

I examine Carrier's and Ladyman's structural realist ('SR') explanation of the predictive success of phlogiston chemistry. On their account, it succeeds because phlogiston chemists grasped that there is some common unobservable structure of relations underlying combustion, calcification, and respiration. I argue that this SR account depends on assuming the truth of current chemical theory of oxidation and reduction, which provides a better explanation of the success of phlogiston theory than SR provides. I defend an alternative version of inference-to-the-best-explanation scientific realism which I call 'Best Current Theory Realism' (BCTR) and argue that it can answer the pessimistic meta-induction.

Does Structural Realism Provide the Best Explanation of the Predictive Success of Science?

(1) Introduction

Scientific realists are committed to the view that some scientific theories—those which exhibit an appropriate degree of empirical success—make claims about unobservable phenomena that are approximately true. Many hold that inference-to-the-best-explanation can confirm the realist view, in conjunction with the claim that the best explanation of the success of scientific theories is the realist view that they, or some components of them, are true. The central difficulty with the explanationist argument is the existence of many scientific theories which were successful in their time but are now rejected as false in light if the emergence of better theories (Laudan, 1981, 1984). Theory change seems to provide counter-evidence to the realist claim that the truth of successful theories is what best explains their success.

Realists have responded by tightening the criteria of successful theories (e.g. requiring 'novel' predictions), restricting the claim of truth to components of theories essential to their success and defending a continuity of reference across change in the ontological posits of theories (Psillos, 1999). But the problems inherent in these realist strategies motivate structural realism (SR) —which limits the realist commitment to the truth of mathematical equations and relations preserved across theory-change, independently of 'theory' claims concerning the unobservable entities and mechanisms referred to in such equations.

My aim is to evaluate the ability of SR to explain the success of theories. I argue that

the plausibility of SR depends on assuming the truth of our best current theories and their substantive claims concerning unobservable entities and processes. I argue that the plausibility of SR's explanation of the success of superseded theories depends on an antithetical view I call 'Best Current Theory Realism' (or BCTR). BCTR limits the realists' commitment to the truth of our best current theories, and thus rejects the continuity-of-true-components hypothesis central to SR, and standard versions of realism. I defend BCTR as offering the best explanation of both (1) the success of superseded and 'falsified' theories and (2) the success of our best current theories.

I evaluate SR's ability to explain success through an examination of Ladyman's and Carrier's SR account(s) of the success of phlogiston chemistry. But we set the stage by starting with Worrall's well-known SR account of Fresnel's theory of light and some criticisms of it advanced by Psillos. My aim is to clarify what SR needs to explain in order to 'explain success'.

(2) Fresnel and the Motivation for SR

Fresnel's account of the propagation of light in an ethereal medium was successful in predicting and explaining a wide range of observed features of the diffusion of light (Worrall, 1989, 1989c, 1990, 1994). It made novel predictions such as the observation of an antecedently unlikely white spot in the center of the shadow of an illuminated circular screen. The theory qualifies as a genuine empirical success, on stringent realist criteria. Yet its claims concerning the nature of light are false. Its explanations and predictions rest on the false claims that a luminiferous ether of molecules is the medium that carries light waves;

that the amplitude of light waves correlates with the velocity of the displacements of ether molecules; that the transverse vibrations of light rays is proportional to the oscillations of ether molecules. All of these hypotheses are abandoned by Maxwell's theory of the electromagnetic field. This situation motivates the realist's desire to find another account of Fresnel's success, one that identifies true components of the theory that do not involve its ontological claims about unobservable entities and processes. The realist's continuity hypothesis requires that such true components are preserved by Maxwell's theory. For SR, the solution is the mathematical structure of relations captured by the equations of Fresnel's theory and preserved by Maxwell, whose theory of the electromagnetic field extends Fresnel's equations to describe electric and magnetic phenomena, in addition to the propagation of light. So although Fresnel's substantive theoretical claims about light are rejected by Maxwell's theory, "there is nonetheless a structural, mathematical continuity between the two theories" (Worrall, 1990a, 21). Thus the best explanation of the success of Fresnel's theory, and that of Maxwell's as well, is that the equations of both are accurate representations of the unobservable structure of relations underlying electromagnetic phenomena.

SR strips down a theory's source of success to its 'uninterpreted' equations. Critics wonder whether such stripped-down mathematical equations can explain theories' empirical success. Some philosophers have simply assumed that a theory's empirical success includes its explanatory power, not just its predictive success. It is doubtful that SR's bare-bones equations can explain a theory's empirical success, because its explanatory power depends on its substantive ontological claims concerning the unobservable entities and processes. The ability of Fresnel's theory to explain phenomena involving the diffusion of light depends

on the substantive hypothesis that a luminiferous ether of molecules is the carrier of light waves. Clearly, it would be a considerable epistemic virtue of IBE realism if it could account for theories' explanatory success. Nonetheless, SR can hold that IBE realism is well enough confirmed if it can explain theories' predictive success.

For these reasons, I will assume that IBE realism, and thus SR, is confirmed if it can provide the best explanation of theories' predictive success. So the issue is whether a theory's bare-bones uninterpreted equations can explain its predictive success. Psillos has given a provocative argument in the negative (Psillos 1999, 153-159). Against SR, a theory's uninterpreted equations alone do not explain its predictive success, because other components of the theory are required by its power to yield any predictions whatsoever. For example, the predictions provided by Fresnel's theory of light require substantive theoretical claims about the conservation of energy, the geometric arrangement of light rays where two media meet, the relation of the amplitude of light waves to the velocity of the displacement of ether molecules, etc. A theory's predictive success will depend on its substantive hypotheses concerning unobservables, background knowledge, auxiliary assumptions, bridge laws, etc. If this is so, IBE realism's attempt to explain theories' predictive success will also founder on the problem of theory change. The uninterpreted equations on which SR relies will not suffice to explain theories' predictive success.

But Psillos' challenge rests on the assumption that an explanation of a theory's predictive success must provide an account of its power to make the predictions that succeed. SR can reject this assumption and insist that it only needs to explain why theories' predictions succeed—that is, come true. The truth of the equations of a theory, by themselves, may explain why its predictions succeed, which is all SR will need. For SR, a

theory's predictions succeed because its equations accurately represent the structural relations between unobservable entities, quite apart from the way these entities or their properties are identified and characterized by the theory. A theory is successful because it gets something right concerning the structure of relations in nature causally responsible for the phenomena that the theory predicts.

This is a perfectly intelligible SR-based explanation of theories' predictive success. But is it a plausible explanation, or the best explanation? Skeptics may worry whether equations can be true apart from the entities and processes they describe, or whether relations in themselves have causal powers, independently of the entities and mechanisms that bear these relations. Skeptics may worry about the whole distinction between descriptions of entities or their properties, on the one hand, and descriptions of their 'relations' on the other. I will circumvent these issues which are murky. Rather, I will examine the structural realist accounts of phlogiston theory's success provided by Carrier and Ladyman. I will argue that the plausibility of their accounts tacitly depends on assuming the truth of current chemical theory and that this claim opens the way onto a better explanation of the success of theories which I call 'Best Current Theory Realism', or BCTR. But why phlogiston theory? Does it provide a useful or telling 'test-case' for SR?

(3) The Phlogiston Theory as a Test-Case for SR

The example of Fresnel's theory of light and Maxwell's electromagnetic theory provides a paradigm of SR because mathematical structure is so obviously retained in this case of theory-change and clearly has a central role in their empirical success. This case

raises the issue of whether SR can provide plausible accounts of cases where mathematical equations are not central to theories' success. Phlogiston theory is one such case. There are several first-rate treatments of this case (Musgrave 1976, Pyle 2000). These case-studies provide strong evidence, marshaled by Ladyman and Carrier, that phlogiston theory enjoyed substantial empirical success on stringent realist criteria like 'novel predictions' (Carrier 2004, Ladyman 2008). So, realists should be able to handle it. No scientific realist holds that the term 'phlogiston' genuinely refers to anything, or that the theory's claims about it are true, despite its substantial empirical success. Thus is provides an excellent opportunity for SR to show that it can succeed, despite the problem of theory-change, and for a case in which equations do not do the work required by its notion of structure. SR does not stand or fall on the basis of its account of phlogiston chemistry, given the large number of theories it may handle better than rival versions of realism. Nonetheless, I argue that the way Carrier and Ladyman explain success in this case may be symptomatic of some general weaknesses in SR itself, and motivate the rival view BCTR.

The empirical successes of phlogiston chemists are persuasively set out by Ladyman and Carrier. These chemists [e..g Becher, Stahl, Priestly, Scheel, among others] demonstrated the existence of empirical regularities concerning the process of combustion, calcination, and respiration, and the effects of these processes [now known as oxidation and reduction] on the properties and weight of wood, calxes, metals, and other substances. Furthermore, their hypotheses concerning the attributes and causal powers of phlogiston generated a unifying explanatory and predictive account of these phenomena. Thus the theory could account for combustion and calcination as the release of phlogiston from the objects into the air, generating the 'phlogistication' of the air and the 'dephlogistication' of

the objects. Both sorts of observable processes could be understood as the result of the behavior of phlogiston. Furthermore, the theory provided a unifying account of other salient phenomena. Why do calxes and metals exhibit different sensible properties? Because all metals contain phlogiston, which generate their metallic properties, while calxes lack phlogiston. Why do substances such as wood and coal end up weighing less as the result of combustion? Because they lose phlogiston. When objects undergo combustion in a confined space, why does the combustion terminate more rapidly in the presence of animals than it does in the presence of plants? Because animal respiration fills the surrounding air with phlogiston, inhibiting the release of phlogiston from an object undergoing combustion, and thus the combustion itself. Plants, on the other hand, absorb phlogiston from the air, generating a more favorable environment for combustion, the release of an object's phlogiston into the surrounding air. Phlogiston theorists also made "novel" predictions, employing the theory to both predict and explain 'new' phenomena—either unknown or ignored at the point when the theory is elaborated to accommodate its central problems and phenomena. Scheel used the theory to correctly predict that new acids (e.g. formic acid, lactic acid, etc.) would be discovered down the road [Ladyman 2008]. Priestly accurately predicted that pure metals would result from heating certain calxs in inflammable ('phlogisticated') air [Carrier 2004].

On the account given by Ladyman and Carrier, phlogiston chemists owed their predictive success to the fact that they got something importantly right about the structure of chemical reactions—namely that there is a common unobservable structure of relations underlying combustion, calcification, and respiration, making them the same kind of process. In current chemistry, this structure is identified as the inverse processes of oxidation and

reduction. Thus phlogiston theory succeeded in unifying three different classes of observable phenomena as the result of the same unobservable structure of relations. As Ladyman puts it, phlogiston theory "captured one great truth retained by Lavoisier in his oxygen theory, namely that combustion, respiration, and calcification are all the same kind of reaction (viz. 'oxidation') and that these reactions have an inverse, namely reduction" (Ladyman, 2008). In Carrier's terms, SR is committed to a "natural kind realism" because it explains the strong success of theories, such as the phlogiston case, as a result of the fact that they posit some unobservable mechanism(s) which show that apparently different sorts of phenomena are really "equal in kind" (Carrier 2004).

This account of the case contains a powerful insight into the achievements of phlogiston chemistry. The issue is whether this account supports SR. To begin with, notice that the evident plausibility of Ladyman's formulation of the 'one great truth' discovered by phlogiston theorists, and preserved by Lavoisier, assumes the truth of current chemistry's claims concerning oxidation and reduction. We know combustion, respiration, and calcification are all the same kind of reaction on the basis of post-Lavoisier chemical knowledge of the nature of oxidation and reduction, and of the entities and mechanisms involved in these processes. But structural realists cannot avail themselves of the substantive ontological claims of any theory, including our best current theories—without abandoning SR in favor of BCTR.

Can we formulate the 'one great truth' uncovered by phlogiston theory, responsible for its predictive success, without assuming our current knowledge of oxidation and reduction? Phlogiston chemists were committed to a relation of unity or sameness between three kinds of *observable* phenomena. Perhaps this conviction constitutes the one great truth

responsible for the success of the theory and preserved in its successors. What SR needs, however, is some *unobservable* and *underlying* structure of relations which is supposed to explain the predictive success of phlogiston theory. What is it? What truths describe it? The truth 'that combustion, respiration, and calcification are all the same kind of reaction' does not describe any underlying structure of relations, so how can it provide an explanation of anything? The case is different for mathematical theories like that of Fresnel and Maxwell, where SR can appeal to their equations to describe the structural relations and true components that do the work for realism.

(4) Motivating the Move from SR to BCTR

This brings us to the nub of the issue. SR can hold that the one great truth discovered by phlogiston theory is simply that there is *some* unobservable structure of relations underlying combustion, respiration, and calcination which make them the same kind of reaction. The question is whether this true component of the theory provides an adequate explanation of its predictive success. It seems like a weak explanation, but that poses no problem for SR if it is the only and thus the best realist explanation that is not vulnerable to the problem of theory-change. My argument is that it does not provide the best realist explanation. If we avail ourselves of the insights of post-Lavoisier chemical theory, the result is a much better realist explanation of the success of phlogiston chemistry than the 'stand-alone' structural account; though admittedly an explanation based on current chemical knowledge cannot be better if it succumbs to the problem of theory-change and the pessimistic meta-induction (which is addressed below).

Why does current chemical theory provide a better realist explanation? The SR account explains the predictive success of phlogiston theory as the result of its insight that there is some unobservable structure of relations responsible for all three sets of observable reactions. But it is current chemical knowledge of oxidation and reduction which provides compelling evidence that there is indeed such an unobservable structure of relations, correctly identifies the structure, and explains how it generates and unifies the phenomena of combustion, respiration, and calcination.

Scientific realists typically hold that the success of a superseded theory in a field is best explained by components of the theory that are preserved by the more successful current theories in that field. My claim is that the plausibility of this backward-looking realist strategy tacitly depends on using our best current theories to identify the truthful components of superseded theories. Furthermore the realist conviction that these truthful components can explain the superseded theory's success derives from its plausibility from the fact that our best current theories—with their full range of hypotheses concerning unobservable entities and processes—also employ some modified version of these components in achieving the greater empirical success of our current theories. To illustrate my argument, consider the SR conviction that the predictive success of Fresnel's theory of light is explained by the accuracy of its equations, the fact that they capture the structure of relations underlying the propagation of light. The explanation only works with the benefit of hindsight. The triumph of Maxwell's theory of the electromagnetic field utilizes these equations to account for the phenomena of light, electricity, and magnetism successfully treated as field dynamics. We thus rely on the truth of our best current theory to identify what Fresnel got right and to fix its role in generating the phenomena he successfully predicted. So, my argument is that the

truth of Maxwell's theory provides the best explanation of Fresnel's success. It is this explanation which is required to confirm the role of his equations in that success, a role dependent on knowledge of the electromagnetic field.

For the sake of argument, let us assume that our best current theories provide the best explanation of the success of superseded and 'falsified' theories. The result is a form of inference-to-the-best-explanation realism, or BCTR. BCTR departs in a sharp way from SR, and more standard forms of realism, by rejecting the need to provide an explanation of the empirical success of every theory solely in terms of its truthful components (structural or otherwise). Realists may identify this move to break the inferential connection between the success of theories and their truth, or true components, as the abandonment of scientific realism itself. But BCTR does not break the inferential connection between success and truth. Rather BCTR reinterprets this inferential connection as one which holds between (1) the success of theories in a scientific field and (2) the truth of our best current theories in that field on the ground that (2) provides the best explanation of (1).

But SR has a powerful rejoinder to my argument for BCTR so far. SR defends its structural account of the success of theories as the best realist explanation of success which does not fall prey to the problem of theory-change and the pessimistic meta-induction to the conclusion that all theories are probably false—at least in their claims concerning unobservable entities and processes. If BCTR is undermined by the problem of theory-change then it does not provide the best explanation of successful theories, and SR may claim that title.

(5) Is BCTR Undone by the Problem of Theory-Change?

The problem of theory-change is often taken to support a pessimistic meta-induction to the probable falsity of all scientific theories. But there is a paradox here which lends some support to BCTR. In itself the problem of theory-change starts with the observation that there are many theories in the history of science which were successful but later discovered to be false. This observation is taken to support an inductive inference to the conclusion that in all likelihood, our most successful current theories are also false. The paradox arises from the fact that the premise of the pessimistic induction—the existence of many successful-but-false theories—depends on the assumption that our best current theories are true, which contradicts the conclusion we are supposed to draw from the inductive inference, or that these current theories are most likely false. Without the assumption that our best current theories are true, there would be no ground for taking successful-but-superseded to be false. They are falsified by subsequent and current theories, on the assumption that the latter are true! But this is precisely the realist claim concerning our best current theories defended by BCTR. Without this realist claim, the only inductive argument that remains is one from the fact that past theories were successful but rejected or superseded, to the likelihood that our most successful current theories will also be rejected. But mere theory-change does not bear on the truth or falsity of any of these theories or on the claims of scientific realism. Indeed, even the conclusion of the pessimistic induction that our best current theories are probably false depends on a realist hypothesis that they will be falsified by new more successful theories that we will know to be true. Ironically, the pessimistic induction turns into an optimistic induction concerning the future emergence of true theories.

Nevertheless, BCTR requires an independent defense of its realist claim that our best

current theories are true and it is reasonable to regard them as such. As a form of inference-to-the-best-explanation realism, BCTR needs to establish that the truth of our best current theories provides the best explanation of the predictive success of science. In the above critique of SR, I have argued that taking our best current theories to be true yields the best explanation of the success of their superseded or 'falsified' predecessors, and in any case, a better explanation than the 'stand-alone' structural explanation provided by SR. So BCTR preserves the realist's desired inferential connection between success and truth, but reinterprets the connection as one that holds between the success of superseded theories in a field and the truth of the best current theory in that field. This inferential connection provides some confirmation for BCTR.

The confirmation of BCTR can be strengthened if it can be shown that the truth of the best current theories also provides the best explanation for *their* success, as well as that of their predecessors. But at this point, the problem of theory-change raises its ugly head once again to challenge BCTR in the very manner that SR escapes, which may make it superior to BCTR. The fact that many theories were successful but false (in their claims concerning unobservable entities and processes) seems to undermine the inferential connection between success and truth. Why should the success of the best current theories be any different, calling into question the inference from their success to their truth, and an appeal to their truth, to explain their success. The answer given so far in this essay is meant to support the inference from the success of superseded theories in a field to the truth of the best current theories in that field—on the ground that it provides the best explanation of the success of their predecessors. But can it also be shown that the truth of our best current theories provides the best explanation of their success as well, strengthening the case for BCTR?

Why should it, if superseded theories were successful but false?

The key to an answer is to identify a property of the best current theories lacking in their predecessors and justifying the realist explanation. Our best current theories enjoy a measure and quality of predictive success unique in the history of the whole scientific field. They are unique in that they alone realize the highest standards of empirical success and confirmation in the field and thus often raise its standards of accuracy, scope, consilience, completeness, unification, and simplicity. This is a fact about the best current theories which distinguishes them from their predecessors. If the question is posed of why the best current theories stand at the apex of empirical success, the hypothesis that it is because they are true provides a plausible— and perhaps the best—explanation.

On this basis, BCTR also has the resources to break the pessimistic meta-induction. For if (1) our best current theories attain a unique empirical success absent in (2) their predecessors, this difference between (1) and (2) undermines the inference from the falsity of (2) to the likelihood that (1) are also false. Yet defenders of the pessimistic meta-induction will not be convinced that the best current theories possess the unique property of attaining a distinctive empirical success at the pinnacle of confirmational virtues. They may argue that this fails to be a unique property because superseded theories also exhibited this property in their time and place. So the scientific realist would have been committed to the truth of these theories at the time and in the context where they stood at the height of empirical success.

True enough, but this move fails to save the pessimistic meta-induction because it misconstrues the status of scientific realism as an empirical hypothesis bound to explain the body of evidence available to us. For BCTR, the fact to be explained is that our best current theories succeed in realizing the most demanding standards of success in the whole history of

their fields—and their predecessors do not: This is the evidence which scientific realism wants to explain. If, as I argue, the best explanation of this fact is (1) that these current theories are true while (2) their predecessors are false, then this inference-to-the-best-explanation trumps the inductive inference from the falsity of (2) to the falsity of (1). Bringing the argument together, the pessimistic induction is undermined if the truth of our best current theories provides the best explanation of their success, and also that of their falsified predecessors.

But there is yet another challenge to BCTR posed by the problem of theory-change, and induction from the fate of theories in science. Successful theories are generally superseded by more successful ones, making it highly probable that our best current theories will sooner or later be displaced by others that satisfy yet higher standards of predictive success. Call this 'the optimistic meta-induction'. Does this induction cast doubt on BCTR and its commitment to the truth of our best current theories? BCTR, as an empirical hypothesis, must rest on available evidence and is fallible in light of new evidence. If and when our best current theories give way to yet more successful theories, this would not refute BCTR, but it would transform what the realist has to explain, and which theories are taken to be true. BCTR is committed to the truth of our best theories because that provides the best explanation of the success of science. Given our present evidence, our best current theories are the best theories and ground a realist commitment to them. This is all that BCTR requires, and is thus not threatened by the possibility or likelihood that what count as the best theories will alter in time. If and when more successful theories arise, then the realist commitment to their truth may provide the best explanation of the success of science. In this way, BCTR saves scientific realism from pessimistic and optimistic inductions concerning

the future of science. For realists, inference-to-the-best-explanation trumps inductive inference in the determination of which scientific theories it is reasonable to regard as true!

My conclusion is that BCTR is not undone by the problem of theory-change, and may thus stand as a better explanation of the predictive success of science than SR.

References

1. Carrier, M. (2004). "Experimental Success and the Revelation of Reality: The Miracle

   Argument for Scientific Realism." (In M. Carrier, J. Roggenhofer, G. Küppers, and Ph. Blanchard (Eds.), *Knowledge and the World: Challenges Beyond the Science Wars*. Berlin: Heidelberg, and New York: Springen-Verlag.)
2. Ladyman, J. (2008). "Structural Realism versus Standard Scientific Realism: The Case of Phlogiston and Dephlogisticated Air," forthcoming in Synthese 2009; *Theoretical Frameworks and Empirical Underdetermination Workshop*. University of Düsseldorf, April 2008: University of Pittsburgh [Electronic version].
3. Laudan, L. (1981). "A Confutation of Convergent Realism," *Philosophy of Science* 48: 19-49.
4. Laudan, L. (1984). "Explaning the Success of Science." (In J. Cushing *et al.* (Eds.) *Science and Reality*. Notre Dame: Notre Dame University Press.)
5. Musgrave, A. (1976). "Why did Oxygen Supplant Phlogiston? Research Programmes in the Chemical Revolution". (In C. Howson (Ed.) *Method and Appraisal in the Physical Sciences*. Cambridge and New York: Cambridge University Press.)
6. Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. (London: Routledge)
7. Pyle, A. (2000). "The Rationality of the Chemical Revolution". (In R. Nola and H. Sankey (Eds.) *After Popper, Kuhn, and Feyeraband: Recent Issues in Theories of Scientific Method*. Dordrecht, Boston, and London: Kluwer University Press.)
8. Worrall, J. (1989). "Structural Realism: The Best of Both Worlds?" *Dialectica* 43:

99-124.

9. Worrall, J. (1989c). “Fresnel, Poisson, and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories”. (In G. Gooding *et al.* (eds) *The Uses of Experiment*. Cambridge: Cambridge University Press.)
10. Worrall, J. (1990). “Scientific Realism and the Luminiferous Ether: Resisting the ‘Pessimistic Meta-Induction’”. Unpublished manuscript.
11. Worrall, J. (1990a). “Scientific Revolutions and Scientific Rationality: The Case of the ‘Elderly Holdout’”. (In C. W. Savage (Ed.) *Scientific Theories*, Minnesota Studies in the Philosophy of Science, Vol. 14. Minneapolis: University of Minnesota Press.)
12. Worrall, J. (1994). “How to Remain (Reasonably) Optimistic: Scientific Realism and the ‘Luminiferous Ether’”. (In D. Hull, M. Forbes, and R. M. Burian (Eds.) *PSA 1994*, Vol. 1. East Lansing, MI: Philosophy of Science Association.)

The Value of Cognitive Values
Heather Douglas
To be presented at PSA 2012 and published in the proceedings

Word Count: 4914

Abstract (100 words):
Traditionally, the cognitive values have been thought to be a collective pool of considerations in science that frequently trade against each other. I argue here that a finer grained account of the value of cognitive values can help reduce such tensions. I separate the values into three groups, minimal epistemic criteria, pragmatic considerations, and genuine epistemic assurance, based in part on the distinction between values that describe theories *per se* and values that describe theory-evidence relationships. This allows us to clarify why these values are central to science and what role they should play, while reducing the tensions among them.

Introduction

The value of cognitive values (also called theoretical virtues or epistemic values) has been underdeveloped in philosophy of science. They have largely been considered together in one group, and when examined in this light, they seem to trade off against one another, creating as much tension as guidance for scientific inference. Although some work has examined a particular value in greater depth and attempted to ground a justification for its importance in an epistemic argument (e.g. Forster & Sober 1994), for the most part, the values have been justified collectively and historically, i.e., that some set of values is (by and large) what has been important to scientists in their practice, and that that should be good enough for philosophers of science (e.g., Kuhn 1977).

This paper will attempt a more robust justification. Through the tactic of organizing the conceptual terrain of cognitive values, I will argue that there are at least three distinct groups of values that normally get lumped together. Once the values are divided into these groups, it is clearer why the values are important and what their value to science and to scientists is. Justifications, clarifying the value of cognitive values, then follow. Creating these divisions requires finer grained appraisals of the values than has been customary. For example, internal consistency will be considered distinct from external consistency. Simplicity has two distinct aspects as well, as does scope. This paper does not make the claim that the terrain mapped here provides a complete account of these values, but the kind of complexity presented can be a starting point for further discussions and amendments.

Another benefit of clarifying the terrain is that the supposed tensions among the values prove to be far less common and problematic than is often presumed. Once the bases for the values becomes clearer, their functions in science become clearer, and thus which should be important when is clarified. In addition, as we will see, the values within a group are shown to often pull together rather than against each other.

Finally, organizing the terrain and mapping the value of cognitive values will also enable us to address the criticisms raised concerning the canonical distinction between epistemic/cognitive and non-epistemic/non-cognitive values (e.g. Rooney 1992) and criticisms over what should count as a cognitive/epistemic value (e.g. Longino 1996).

First, I will provide a brief overview of how the standard view on cognitive values developed. Then, I will offer a more nuanced terrain for those values than has been traditionally offered. I will proceed to show how both tensions among the values are reduced (albeit not eliminated) and how the justifications for the various values are clarified. Finally, I will draw implications from this re-organization of the terrain.

A Brief History of Cognitive Values

Philosophers of science have long referred to and discussed various qualities of scientific claims deemed important in science. In the 20th century, philosophers such as Duhem (e.g., 1906, 171, 217), Popper (e.g., 1935, 61-73, 122-128) and Levi (1960, 354; 1962, 49) famously described a range of qualities (and sometimes provided reasons for the importance of those qualities). But it was not until Kuhn's 1977 paper that these qualities became widely known as values, and the discussion was framed in terms of values internal to science. For Kuhn (1977), McMullin (1983), Laudan (1984), and Lacey (1999), the values were a collective (if evolving) set. And there were clear tensions and tradeoffs among the various values or virtues thought relevant at any given time. One might gain scope in a theory, but lose precision. One might gain simplicity, but lose scope. Understanding the history of science meant understanding how scientists made those trade-offs (or shifted their interpretation of those values) in the course of scientific debate.

But the collective pool of these values turns into a problematic swamp when one attempts to find a grounding for the values. This problem was worsened by the tendency of philosophers, in an attempt to make the values appear less overwhelming, to collapse various attributes together. Thus, although some distinguished internal consistency (minimal logical consistency of a theory) from external consistency (broader considerations of whether a theory fit with prevailing scientific views), other philosophers collapsed the two, and considered consistency *tout court* (e.g., Kuhn 1977, 357 vs. McMullin 1983, 15) This makes it harder to see how to justify consistency. While internal consistency can be viewed as a minimal requirement of empiricism (Duhem 1906, 220; Popper 1935, 72), external consistency is nothing of the sort, and is valuable only insofar as one's confidence in the rest of scientific theory is high. Or consider how explanatory power can be viewed either as an ability of a theory to elucidate particular pieces of evidence with great detail or as an ability of a theory to bring under one conceptual umbrella multiple disparate areas (which can also be conflated with scope). Both are clearly valuable, but for quite different purposes and reasons.

It is time to extricate ourselves from this swamp. Laudan (2004) made the first steps in this direction when he divided theoretical virtues into those that were genuinely epistemic

(truth indicative) and those that were cognitive (valued by scientists for other reasons). He suggested that few of the traditional theoretical virtues (construed as the swampy collective described above) have genuine epistemic (that is, truth-indicative) merit. Two that did (on his view) were internal consistency and empirical adequacy. Laudan's distinction is a good start on the problem, but I will go further here, dividing up the terrain of cognitive values further in an attempt to elucidate their strengths, their purposes, and their justifications.

The Terrain of Cognitive Values

Two distinctions will help further our project. First, following both Laudan (2004) and Douglas (2009), we can distinguish between ideal desiderata and minimal criteria. We might prefer one grand, simple, unified theory of great scope that explains everything, but in practice we are willing to settle for less. (Indeed, some arguments for pluralism suggest we should be happy with a complex plurality of perspectives. See, e.g., Kellert, Longino & Waters 2006; Mitchell 2009.) In contrast, there are some virtues or values that any acceptable scientific theory must instantiate (e.g. internal consistency). We might accept a theory that falls short on these criteria out of shear desperation, but we would know something was wrong and work furiously to correct it.

Second, it is important to note that in discussing the set of cognitive values, philosophers have lumped together two different kinds of things in science to which cognitive values can apply. By "apply", I mean that which the values are thought to describe, or the object of instantiation for the value (i.e., what has the value). The object of instantiation can either be a theory *per se* or the theory in relation to the evidence thought to be relevant to it. There are thus two different directions for assessment when using cognitive values: are we describing the theory itself or the theory in relation to the available evidence?

To see how crucial these two different targets for cognitive values can be, consider the value of scope. If we are talking about a theory with scope (and just the theory), the theory might have the potential to apply to lots of different terrain or to wide swaths of the natural world (i.e. the claims it makes are of broad scope), but whether it in fact does so successfully can still be up in the air. Any proposed grand unified theory can be considered to have scope in this sense—it has broad scope, but not in relation to any actual evidence yet gathered under that scope. Contrast that with a theory that already does explain a wide range of evidence and phenomena—so that the scope applies to a theory in relation to broadly based evidence (e.g. evidence from different phenomena or evidence gathered in different ways). Here the value of the cognitive value is quite different, and brings with it an epistemic assurance from the diversity of evidence supporting the theory.

A similar point can be made with regards to simplicity. A simple theory (that is, just a simple theory, and not where simplicity is describing a relation to evidence) might be *prima facia* attractive, but unless we think the world actually is simple, we have little reason to think it true. A simpler theory, all other things being equal, is not more likely to be true. Contrast this with a theory that is simple with respect to the complex and

diverse evidence that it captures. The simpler theory, in relation to the evidence it explains, is more likely to not be overfit to the evidence and thus more likely to be predictively accurate. (Forster & Sober 1994) In such a case, simplicity has genuine epistemic import.

With these two distinctions in mind—1) what we want our values for (minimal criteria vs. ideal desiderata) and 2) to what the value applies (the theory *per se* vs. the theory with respect to evidence)—we can turn to the terrain for such values. There are three groups into which we can divide the cognitive value terrain:

Group 1: Values that are minimal criteria for adequate science

There are values that are genuinely truth assuring, in the minimal sense that their absence indicates a clear epistemic problem. If a claim or theory lacks these values, we know that something is wrong with our empirical claim. Thus, these are truly minimal criteria, values that must be present if we are to be assured we are on the right track. These values include internal consistency (which is about the theory *per se*) and empirical adequacy (as measured against existing evidence, not all possible evidence, and thus is about the theory with respect to evidence). Philosophers as diverse as Duhem (1906), Popper (1935), Laudan (2004), and Douglas (2009) have noted these values as minimal criteria. This group could be divided along the lines of Group 2 and 3 below using the second distinction (regarding the instantiation of the value), but because it is so small, I leave them together here. Because both of these minimal criteria have clear epistemic import (theories failing these criteria are not good candidates for our beliefs), keeping them in the same group helps clarify their function.

Group 2: Values that are desiderata when applied to theories alone

There are values that, when instantiated solely by the theory or claim of interest, give no assurance as to whether the claims which instantiate them are true, but give us assurance that we are more likely to hone in on the truth with the presence of these values than in their absence. As such, these might be considered strategic or pragmatic values. Douglas (2009) emphasizes the term *cognitive* values, as an aid to thinking; Dan Steel has called them extrinsic epistemic values (2010). These include scope, simplicity, and (potential) explanatory power. When theories (or explanations or hypotheses) instantiate these values, they are easier to work with. Simpler claims are easier to follow through to their implications. Broadly scoped claims have more arenas (and more diverse areas) of application to see whether they hold. Theories with potential explanatory power have a wide range of possible evidential relations. (I say potential because if the theory has actual, known explanatory power, that implies that evidence is already gathered under its umbrella and this would bring us to the next category of values.) It is easier to find flaws in the claims and theories that instantiate these values. It is easier to gather potentially challenging (and thus potentially strongly supporting) evidence for them. In this sense, all of these values fall under the rubric of the fruitfulness of the theory.

Group 3: Values that are desiderata when applied to theories in relation to evidence

Finally, we should consider values that might sound similar to pragmatic cognitive values (group 2), but because they qualify the relationship between theory and evidence, rather than just theory itself, they provide a different kind of assurance. Whereas group 1 assured us that we have a viable scientific theory (genuine epistemic assurance), and group 2 assured us that if we were on the wrong track, we should find out sooner than otherwise, group 3 provides a particular kind of genuine epistemic assurance. It provides assurance against ad hocery, and thus assures us that we are not making a particular kind of mistake. One of our most central concerns in science is that we have made up a theory that looks good for a particular area, but all we have done is make something that fits a narrow range of evidence. If our theories are *ad hoc* in this way, they will have little long term reliability or traction moving forward. Instantiation of these values in the relation between the theory and the evidence that supports it provides assurance that we have not just made something up. If a diverse range of evidence can be explained, or the theory fits well with other areas of science (and, crucially, the evidence that supports them), or the theory makes successful novel predictions, we gain precisely the assurance we need. For this reason, these values have genuine positive epistemic import. These values include unification (in terms of explanatory scope, simplicity, external consistency, and coherence), novel prediction, and, modifying these values with an additional layer, precision. (I discuss this group further below.)

What does this map of the terrain clarify? First, with this map we can see that the values do have justifications independent of scientists' historical reliance on them. We can articulate reasons why a scientist should care about these values and clarify what they are good for. There are clear epistemic reasons (independent of any particular objectives of science at any particular period) for demanding that scientific theories be internally consistent and empirically competent. And there are good epistemic reasons for preferring scientific theories which have a broad range of evidence that support them or that instantiate other values in group 3 (more on this below). Finally, there are good pragmatic reasons for scientists to run with a simpler, broader, or more fruitful theory first (group 2) if one is trying to decide where to put research effort next.

Second, as I will argue below, the idea that the values are in a collective pool and pull against each other is misguided. Having this map makes it clearer what the purposes of the values are, and shows that the tensions among the values are not as acute or problematic as they appear when they considered as a collective pool.

Reducing the Tensions among the Values

There are two possible sources of tensions within the terrain I have mapped above. The first arises from tensions among the groups of values. The second arises from tensions within each group. I will address each of these in turn as I argue that tensions with this map have been reduced, albeit not eliminated.

Among the groups, one reduction in tension should be immediately clear.
Minimal criteria do not (or at least, should not) pull against pragmatic fruitfulness

concerns of group 2 or the epistemic assurance concerns of group 3. Minimal criteria come first, and both must be met. Indeed, one cannot tell whether one has an empirically competent theory without minimal internal consistency. Now, in practice, scientists may still choose to pursue the development of a theory with characteristics of group 2 even in the face of failings in group 1 (minimal criteria). But this must be done with the full acknowledgement that the theory is inadequate as it stands, and that it must be corrected to meet the minimum requirements as quickly as possible. Although philosophers like to quip that every scientific theory is "born falsified," no scientist should be happy about it.

Once the remaining values are divided into the pragmatic cognitive values (instantiated by theories only—group 2) and the epistemic anti-ad hocery assuring values (instantiated by the relations between theories and evidence—group 3), the two groups have less problematic tension *within* each than has been generally thought.

Consider the possible tensions *within the pragmatic cognitive values—group 2*. Recall that within this group, the values describe theories or claims on their own, independent of the evidence which may or may not support them. In this group, all of these values are ultimately about the fruitfulness of the theory, the ease with which scientists will be able to use the theory in new contexts (not necessarily successfully), to devise new tests for the theory, and thus refine, revise, or if need be overhaul completely, the theory. It is true that some scientists will find scope an easier handle with which to further test a theory, as they will find it more amenable to apply the theory in a new arena to which the broadly scoped theory is applicable, and some scientists will find simplicity an easier handle with which to devise further tests. So some tensions may remain around the issue of what will be fruitful for different scientists. But this need not create any epistemic worries, for three reasons. First, the proof will be in the pudding for fruitfulness, and the pudding is relatively straightforward to assess. If the theory cannot be used to devise additional tests, if the scientists are unable to use the aspects of the theory that instantiate the value they prefer, then the value is of no further use in that case. We will be able to tell readily if the instantiation of a pragmatic-based value in fact proves its worth. Second, because this category of values does not provide direct epistemic warrant, but is instead focused on the pragmatic issue of the fruitfulness of a theory, there is little reason to be concerned about divergent scientific perspectives on these values. None of these pragmatic values provides a reason to accept a theory as well-supported or true or reliable at the moment. Group 2 values are simply not epistemic. Third, social epistemological approaches to science (e.g. Solomon 2001, Longino 2002) have made it quite clear that having diverse efforts in scientific research is a good thing for science. It has been argued that diversity of efforts in science is crucial for the eventual generation of reliable knowledge. So having diverse views about what makes a theory fruitful is likely to be good for science. In sum, the values in this group are pragmatic, they are easily assessable by external criteria (are more new tests being produced?), their diversity supports a diversity of epistemic effort, and yet, they do not have direct epistemic import. Whatever tensions arise here can play out in diverse efforts of scientific practice.

Consider next the possible tensions *within group 3*. Because these values do have genuine epistemic import, tensions among them would be central to the problem of

scientific inference and the epistemic assessment of scientific theories. But when examining these values as instantiated by the relation between theories and the evidence that supports them, there is less tension among these values than might be initially supposed. For example, while simplicity, scope, and explanatory power are often thought to pull against each other when considering theories alone (group 2), they pull together when considering a theory in relation to evidence (group 3). A theory that has broad scope over diverse evidence is also simple with respect to that diverse evidence, unifies that diverse evidence, and has explanatory power over that evidence. Indeed, it is this set of relations that Paul Thagard has formalized under his conception of "coherence." (Thagard 2000) Scientists might disagree over which evidence is more important to unify or explain under a particular rubric, either because of different purposes or because of different views on the reliability of the evidence under consideration. But that is a disagreement over which instantiation of a cognitive value is more important, not a disagreement based on tensions among values.

Yet there are still some tensions in group 3. For example, predictive accuracy (or the value of the novel prediction) might pull against the considerations captured by coherence. And indeed, when faced with such a tension, scientists can legitimately disagree, some scientists finding greater epistemic assurance in the successful novel prediction and other scientists finding greater epistemic assurance in the successful unification of evidence or the explanatory power/coherence of a theory. When we have both together, both successful explanation of the available evidence and a surprising prediction (use novel or temporally novel), we have Whewell's consilience (Fisch 1985), which is perhaps the strongest epistemic assurance we have available to us. When consilience is on the table, it is hard for other theories to compete. But we are not always so lucky. Hence genuine epistemic tension is possible here.

There is an additional qualifier for the value considerations of group 3. Whether we are considering the relation between theory and evidence that is some form of coherence or some form of prediction, the precision or tightness of fit between the theory and evidence also matters. The more precise the explanatory relations between theory and evidence, or the more precise the prediction *and* the evidence that tests it (having just one or the other is not helpful), the more we gain the epistemic assurance of group 3. This assurance is that we have not just made our theories up, that they have some empirical grip on the world—they are fundamentally anti-ad hocery assurance. The more precision we have in the relations between theory and evidence, the more assurance we get. The more successful predictions we have, the more assurance we get. The more coherence or explanatory power over diverse evidence we have, the more assurance we get. Because there are these different sources of this kind of assurance, there will be tensions among them in practice. But hopefully why these tensions arise, and what should be done about them, will be clearer.

So what of tensions *between the values of group 2 and group 3*? These two groups aim at different purposes, and thus any apparent conflict can be managed. It is particularly important to note that group 2, the pragmatic cognitive values, have no bearing on what should be thought of as our best supported scientific knowledge at the moment. Just

because a theory looks fruitful (whether because of its innate simplicity, scope, or potential explanatory power) is no reason to think it more reliable now than any other narrower or more complex theory. If one needs epistemic assurance, particularly for an assessment of our best available knowledge at the moment, group 3 is where one should look (after the requirements of group 1 are met). When one needs to figure out what should be said about the state of knowledge now, pragmatic fruitfulness (group 2) concerns have no bearing. When one wants to justify future research endeavors, such pragmatic concerns are central.

In sum, there are no tensions among the groups: group 1 trumps groups 2 & 3, and groups 2 & 3 have different purposes. Within the groups, there are no tensions within group 1, there are productive tensions within group 2, and there remain some tensions within group 3. Thus, while tensions among values remain, they are much reduced from the traditional view. With a clearer account of the bases for such values, we can see their function more clearly, and thus their purposes.

Implications

In earlier accounts of the theoretical virtues, the tensions among them were thought to explain how scientists at any given moment could rationally disagree with each other—different scientists focused on different virtues. Does my organization of the theoretical virtues dissolve this ready-made explanation for rational disagreement? No-- there are still resources we can draw upon to explain disagreement. So, for example, one can still see a tension between the explanatory scope of a theory (with respect to available evidence—group 3) and the predictive precision of its competitor. Such a tension will likely continually arise in scientific practice. Or, consider the tension between a well-supported theory (with group 3 values supporting it) and an underdeveloped theory (with lots of group 2 values and thus lots of potential). The explanations of divergent choices that we give, scientists being risk-takers with new theories or with staying with the older, more developed theories, still hold in the account given here, but with a sharper understanding of the source of the divergent choices. Indeed, we should help scientists distinguish an epistemic assessment from a pragmatic fruitfulness assessment in their commitments to scientific theories. Finally, one could also use the account of the place of social and ethical values given in Douglas 2009 to show how concerns over the sufficiency of evidence (driven by social or ethical values) could generate rational disagreement among scientists (as Douglas argues ethical values in the assessment of evidential sufficiency is a rational role for those values).

So what has been gained by organizing and explicating the various values of cognitive values? First, we can see more clearly where and why such values are indeed valuable. The justification need no longer rest on the contingency of the history of science (although it is certainly illuminated by the history of science). This allows us to note why these values have seemed so central. Groups 1 & 3 have genuine epistemic import, and thus do not bleed across the epistemic/non-epistemic boundary (although their instantiation depends on the available evidence which does depend on cultural values). The pragmatic group 2 can have clear cultural influences on it. Rooney's concerns

(1992) are thus illuminated. It also allows us to assess proposals for alternative sets of values (e.g., Longino 1996). We can consider alternative values under the groups proposed and see if they assist us in reaching our goals.

Second, we can now address the reference often made to these values in other debates with greater conceptual clarity. For example, when critics of the value of prediction (as opposed to accommodation) (e.g., Harker 2008, Collins 1994) attempt to reduce the value of novel prediction to accommodation plus a theoretical virtue (such as unification or explanatory power), we can see both what might motivate such an attempt (they are drawn to the power of group 3) and why it is misguided (the value of novel prediction can be in tension with the value of unification). Finally, if this is indeed a step forward in the clarity of the terrain, there is perhaps hope for a renewed effort in a qualitative theory of scientific inference. But that work must await another paper.

References

Collins, Robin. 1994. Against the Epistemic Value of Prediction over Accommodation. *Nous* 28,2: 210-224.

Douglas, Heather E. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Duhem, Pierre. 1906/1954/1982. *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Fisch, Menachem. 1985. Whewell's Consilience of Inductions—And Evaluation. *Philosophy of Science* 52: 239-255.

Forster, Malcolm and Elliott Sober. 1994. How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions. *British Journal for the Philosophy of Science* 45: 1-35.

Harker, David. 2008. On the Predilections for Predictions. *British Journal for the Philosophy of Science* 59: 429-453.

Kellert, Stephen, Helen Longino, and Kenneth Waters (eds.). 2006. *Scientific Pluralism*. Minnesota Studies in Philosophy of Science vol. XIX. Minneapolis: University of Minnesota Press.

Kuhn, Thomas. 1977. Objectivity, Value, and Theory Choice. In *The Essential Tension*. Chicago: University of Chicago Press, 320-339.

Lacey, Hugh. 1999. *Is Science Value Free? Values and Scientific Understanding*. New York: Routledge.

Laudan, Larry. 2004. The epistemic, the cognitive, and the social. In P. Machamer & G.

Wolters (Eds.) *Science, Values, and Objectivity*. Pittsburgh: University of Pittsburgh Press, 14-23.

Levi, Isaac. 1960. Must the scientist make value judgments? *Journal of Philosophy*, 57, 345-357.

Levi, Isaac. 1962. On the Seriousness of Mistakes. *Philosophy of Science* 29: 47-65.

Longino, Helen. 1996. Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy. In Lynn Hankinson Nelson and Jack Nelson, (eds.) *Feminism, Science, and the Philosophy of Science*. Dordrecht: Kluwer, 39-58.

Longino, Helen. 2002. *The Fate of Knowledge*. Princeton: Princeton University Press.

McMullin, Ernan. 1983. Values in Science. In Peter D. Asquith and Thomas Nickles, (ed.), *Proceedings of the 1982 Biennial Meeting of the Philosophy of Science Association, Volume 1*. East Lansing: Philosophy of Science Association, 3-28.

Mitchell, Sandra. 2009. *Unsimple Truths*. Chicago: University of Chicago Press.

Popper, Karl. 1935/1959/1992/2002. *The Logic of Scientific Discovery*. London: Routledge.

Solomon, Miriam. 2001. *Social Empiricism*. Cambridge, MA: MIT Press.

Steel, Daniel. 2010. Epistemic Values and the Argument from Inductive Risk. *Philosophy of Science* 77: 14-34.

Rooney, Phyllis. 1992. On Values in Science: Is the Epistemic/Non-Epistemic Distinction Useful? In David Hull, Micky Forbes, and Kathleen Okruhlik (ed.) *Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association, Volume 2*. East Lansing: Philosophy of Science Association, 13-22.

Thagard, Paul. 2000. *Coherence in Thought and Action*. Cambridge, MA: MIT Press.

Delayed-Choice Experiments and the Metaphysics of Entanglement

Matthias Egg

February 29, 2012

1 Introduction

Entangled states of quantum systems have played an important role in debates on the metaphysics of contemporary science. Teller (1986), French (1989), Maudlin (1998) and Esfeld (2004) all base metaphysical claims on the quantum mechanical description of entangled two-particle states. Glossing over the substantial differences between these accounts, their general thrust is that quantum mechanics pushes us away from a metaphysics of individual things towards a metaphysics of relations, relations which do not supervene on the intrinsic properties of their relata. Of course, such arguments presuppose that what appears as the description of an entangled state in the quantum formalism actually represents a real relation in the world. This presupposition derives support from the fact that measurements on entangled states exhibit significant correlations, as demonstrated, for example, in experimental tests of Bell's inequality. Furthermore, the flourishing of quantum information theory in the last two decades has resulted in a rather dramatic shift of attitude towards entanglement: While it was once considered a bizarre consequence of the quantum formalism with little importance outside of philosophical debates, entanglement has now come to be recognized as a physical resource which can be experimentally manipulated in various ways.[1]

In the present paper, I will investigate whether this manipulability really supports a realist stance on entanglement relations. More specifically, I will focus on one particular kind of such manipulations, namely *entanglement swapping*. In this process, two pairs (A, B) and (C, D) of entangled particles are created by two independent sources. If one then performs the right kind of joint measurement on particles B and C, the pair (A, D) enters into an entangled state even though A and D have never interacted with each other. Interestingly, this phenomenon has given rise to contradicting ontological conclusions. On the one hand, Clifton (2002, S163) takes it to support a realistic view of entanglement: "It appears that there is sufficient substantiality to entanglement that it can be swapped from one pair of particles to another". On the other hand, Healey (forthcoming, sec. 4) studies an entanglement-swapping experiment in which the measurement

[1] For an extensive review of this development and the current state of play, see Horodecki et al. (2009).

on B and C is performed *after* A and D have been detected. Since this seems to imply that entanglement can be transferred to a pair of particles which no longer exists, Healey concludes:

> The delayed-choice entanglement-swapping experiment reinforces the lesson that quantum states are neither descriptions nor representations of physical reality. In particular, it undermines the idea that ascribing an entangled state to quantum systems is a way of representing some new, non-classical, physical relation between them.[2]

Obviously, the "delayed-choice" clause plays a central role here. I will therefore begin my investigation with a brief reminder of the simple and well-known delayed-choice double-slit experiment, assessing its impact on realism about the state of the quantum system (section 2). A more sophisticated (and more radical) version of delayed choice, the so-called *quantum eraser*, will be discussed in section 3. The case of the quantum eraser is important because it introduces the idea of sorting experimental results into different subensembles, thus raising the question whether these subensembles correspond to real properties of the system. In section 4, I will apply these considerations to the entanglement-swapping experiment and show that if the experiment is carried out in a delayed-choice setting, no actual entanglement swapping occurs. This will, in section 5, lead to the conclusion that delayed-choice entanglement swapping does *not* undermine realism about entanglement relations.

2 Delayed Choice in the Double-Slit Experiment

The double-slit experiment is probably the best known illustration of the basic mystery of quantum mechanics. If quantum particles (e.g., electrons) are sent through a double slit, a characteristic interference pattern appears on the screen behind the two slits. However, this pattern disappears as soon as one tries to detect through which of the two slits each electron passed. It thus seems that the electrons either behave as waves (passing through both slits and producing an interference pattern) or as particles (passing only through one slit and displaying no interference), depending on the kind of experiment we choose to perform. (In the following, I will refer to the two kinds of experimental arrangements as "DS" (for "double slit") and "WW" (for "which way"), respectively.) This is already puzzling enough, but further puzzlement is added by the insight that the decision to perform either a DS or a WW experiment can be taken *after* the electron has passed through the double slit. It was Wheeler (1978) who introduced this idea of a *delayed choice*, and he took it to imply that "the past has no existence except as it is recorded in the present", and that "[t]he universe does not 'exist, out there,' independent of all acts of observation" (41).

[2] Healey advances this argument in the context of his pragmatist approach to quantum theory, which I will not discuss here. Neither will I discuss the positions of those who take quantum information theory to support an epistemic or informational (as opposed to metaphysical) view of the quantum state. See Timpson (2010) for a critique of these approaches.

It is not hard to see how delayed-choice experiments can lead to such anti-realistic conclusions. If we think of the electron as traveling from the source to the double slit and then to the screen where it is detected, a natural question to ask is whether the electron behaved as a wave or as a particle at the time it travelled through the double slit. (That electrons are disposed to behave in either of the two ways is already known from DS and WW experiments without delayed choice.) Now if the type of experiment (DS or WW) is fixed in advance, this determines the behavior of the electron, and a unique story about its wave- or particle-like nature can be told for each type of experimental setup. However, in the delayed-choice case, the experiment-type is *not yet* fixed at the time the electron is at the double slit, so it seems that there is simply no fact of the matter as to whether the electron passes through both slits (as waves do) or through only one slit (as particles do).

It is thus clearly impossible to tell a simple realistic story about what happens at the double slit in a delayed-choice experiment. More sophisticated realistic stories remain, of course, possible, but they do not come without a cost. In the next section, I will argue for such a story, based on the formalism of standard quantum mechanics. I will therefore not have much to say about non-standard quantum theories, such as Bohmian mechanics or the Ghirardi-Rimini-Weber (GRW) theory. But I conclude the present section with some brief remarks about these two theories, in order to illustrate to what extent the delayed-choice double slit complicates the realist's ontological commitments.

At first sight, it seems that Bohmian mechanics has a straightforward answer to the question of what happens at the double slit: Being a particle theory, Bohmian mechanics clearly tells us that each electron goes only through one slit. But it also tells us that the movement of the particle is determined by the wave function, and this raises the tricky question of the ontological status of the latter. Some versions of the theory interpret the wave function as a physical entity which literally guides the particles, or they introduce a so-called quantum potential which gives rise to non-classical forces acting on them. But due to some problems of these interpretations, there is now a tendency among Bohmians to regard the wave function no longer as a physical entity, but merely as a component of the law according to which the particles move (Dürr et al. 1997). However, in their detailed analysis of delayed choice experiments from a Bohmian perspective, Hiley and Callaghan (2006a) manage to avoid a commitment to very bizarre particle trajectories only by relying explicitly on the physical reality of the wave function and the quantum potential, so there is reason to doubt that these entities can really be cut off from the ontology of Bohmian mechanics without a loss.

An alternative way to tell a realistic story about the double slit is given by the GRW theory, which adds spontaneous collapses of the wave function to the Schrödinger evolution. But, as argued by Allori et al. (2008; see also Maudlin 2011, 229-238), this theory is also ambiguous in its ontological commitments. In one version, the wave function describes a *matter density* in space. Again, this seems to suggest a straightforward solution to our problem: The matter density, being a spatially extended field, (almost) always passes through both slits and collapses to a particle-like object only upon interaction with the detecting screen. The fact that the experimental setup can be chosen after the matter wave has passed the double slit then poses no particular problem. However,

according to this story, the result of a WW experiment must be regarded as outright illusory: Even though it *looks as if* the electron went only through one slit (the experiment telling us which one), the fact is that it went through both. An even more severe illusion takes place according to the second version of the GRW theory, which is merely committed to the existence of some events in space-time (the *flashes*), corresponding to the spontaneous collapses of the wave function. In this picture, contrary to what we might take as an unquestionable truth about any double-slit experiment, *nothing at all* travels from the source to the screen.

3 The Quantum Eraser

In the experiments discussed so far, the DS/WW decision is taken after the electron has passed through the double slit, but it obviously has to be taken *before* the electron is detected. Using a *quantum eraser* (Scully and Drühl 1982), even this restriction can be removed. Consider the thought experiment by Scully et al. (1991) depicted in figure 1: In a double-slit experiment with atoms, we place a micromaser cavity in front of each slit. The cavities are designed such that excited atoms passing through them inevitably decay into the ground state by emitting a photon. So for each atom, the corresponding photon emitted in one of the cavities provides us with WW information. However, this information can be "erased" by opening the shutters which separate the two cavities from a thin-film photodetector placed between them. Since this "detector wall" does not discriminate between photons coming from one or the other cavity, the WW information is lost and a DS configuration is reestablished. So the experimenter has two options: He can either leave the shutters in place and detect which of the cavities contains the photon, thereby obtaining WW information, or he can open the shutters, allowing the photon to interact with the detector wall without yielding WW information. Note that he can (in principle) decide between these two options *after* the atom is detected at the screen.

But one might ask how this can really be a choice between a DS and a WW scenario, given that the two scenarios should lead to radically different distributions of atoms on the screen (displaying interference fringes in one case but not in the other). Surely the pattern on the screen can not be changed retroactively? Well, in a certain sense, it can. To see how, a closer look at the quantum mechanical description of the atom-photon-system is necessary.[3] If we denote the photon state by $|1\rangle$ or $|2\rangle$, depending on whether the photon is in cavity 1 or 2, and the spatial wave function of an atom coming through one of the two slits by $\psi_1(x)$ and $\psi_2(x)$, respectively, the state of the system after the atom has passed the slits is

$$|\Psi\rangle(x) = \frac{1}{\sqrt{2}}[|1\rangle\psi_1(x) + |2\rangle\psi_2(x)]. \tag{1}$$

The probability density $\|\Psi(x)\|^2$ associated with this state has vanishing interference terms, because $|1\rangle$ and $|2\rangle$ are orthogonal to each other. Therefore, the distribution of

[3]For the following, I adopt the notation of Englert et al. (1999).

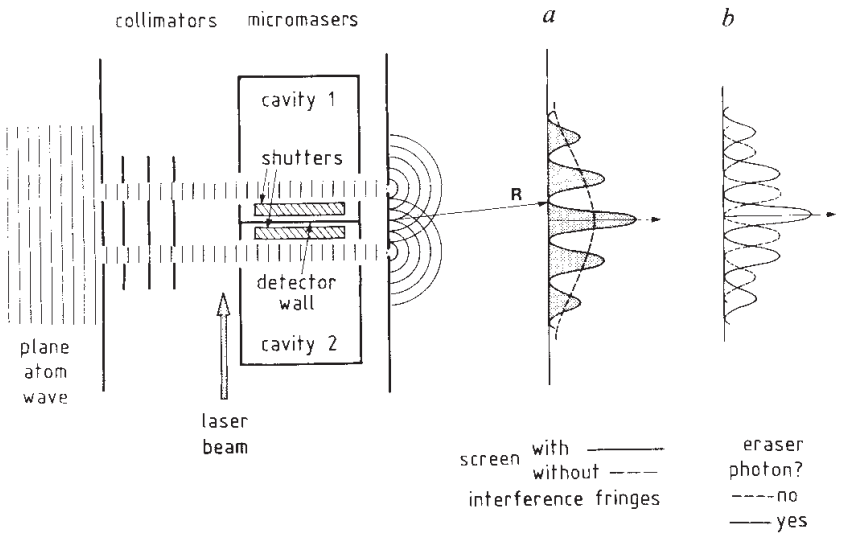


Figure 1: The quantum eraser thought experiment (Scully et al. 1991, 115).

atoms on the screen shows no interference fringes, which is what we expect for a WW experiment. Now let us see what happens if the shutters are opened to erase the WW information. As mentioned above, the detector wall does not discriminate between the $|1\rangle$ and the $|2\rangle$ state. However, it does (maximally) discriminate between the symmetric and the antisymmetric superposition states

$$|+\rangle = \frac{1}{\sqrt{2}}[|1\rangle + |2\rangle] \quad \text{and} \quad |-\rangle = \frac{1}{\sqrt{2}}[|1\rangle - |2\rangle].$$

As a consequence, the detector records only photons in the $|+\rangle$ state and ignores photons in the $|-\rangle$ state. Introducing the corresponding symmetric and antisymmetric states of the atom

$$\psi_{\pm}(x) = \frac{1}{\sqrt{2}}[\psi_1(x) \pm \psi_2(x)],$$

we can rewrite (1) as

$$|\Psi\rangle(x) = \frac{1}{\sqrt{2}}[|+\rangle\psi_+(x) + |-\rangle\psi_-(x)]. \tag{2}$$

Since this is just another way of expressing the same state $|\Psi\rangle(x)$, the probability distribution $\|\Psi(x)\|^2$ is still the one corresponding to the WW setup. But if we restrict our attention to those atoms for which the photodetector records a photon, the contribution from $|-\rangle\psi_-(x)$ vanishes and the probability distribution becomes

$$P_+(x) = |\psi_+(x)|^2 = \frac{1}{2}|\psi_1(x) + \psi_2(x)|^2,$$

which simply corresponds to the result of a DS experiment, displaying the usual interference fringes. Conversely, selecting atoms for which the detector does not record a photon yields $P_{-}(x) = |\psi_{-}(x)|^2$, which corresponds to the complementary "anti-fringe" interference pattern (see figure 1b). So each atom can be assigned to one of the four subensembles "1", "2", "+" and "−" in the following way: By measuring the photons in the WW configuration (shutters closed) we classify the atoms into the subensembles 1 and 2, by measuring the photons in the DS configuration (shutters open), we carry out a $+/-$ classification.

To assess the metaphysical significance of delayed choice in this context, we now need to ask to what physical property (if any) the sorting into subensembles corresponds. Focussing for the moment on subensembles 1 and 2, the answer seems obvious: All atoms in subensemble 1 went through the first slit, the atoms in subensemble 2 through the second one. Englert et al. (1999, 328) endorse this view, but they add an antirealistic twist: "The '... *went* through ...' is not a statement about the atom's past". This reinterpretation of everyday language is motivated by their "minimalistic interpretation" of the quantum state, which I will not further discuss (see footnote 2 above). In a closely related proposal, Mohrhoff (1999) invokes a kind of *retrocausation*, according to which the present determines the past of the atoms. (Notice the similarity to Wheeler's above-mentioned view that the past's existence depends on its being recorded in the present.) This commitment to retrocausation renders Mohrhoff's "reality-of-phenomena point of view" rather unattractive, but one might be willing to accept this consequence. What one should *not* accept is Mohrhoff's claim that "retrocausation is a necessary feature of *any* realistic interpretation of the quantum formalism" (332). He reaches this conclusion by (allegedly) showing that the alternative "reality-of-states point of view" is not viable. Here is his argument:

> Adherents to the reality-of-states view thus find themselves faced with a dilemma. If they ... deny the possibility of retrocausation, they must insist that it is only *as if* the atom had traveled through the first cavity or only *as if* it had been retroactively furnished with a definite phase relation. They cannot say that the atom *really* was in the state ψ_1 (or ψ_2, or ψ_+, or ψ_-, as the case may be). And so they find themselves compelled to foreswear realism and embrace operationalism. And if they stick to realism, they will have to drop the *as if*'s and accept the reality of retrocausation.

But the first thing a reality-of-states view should take seriously is the fact that $|\Psi\rangle$ in (1) and (2) is an *entangled* state of the atom-photon system, so it is clear from the start that none of the four (pure) ψ-states can be ascribed to the atom alone, as long as the system is in state $|\Psi\rangle$. It is true that this commits the realist to Mohrhoff's *as-if*-statements (compare my remarks on the matter-density version of GRW in section 2), but to say that reality differs from what measurements seem to reveal is very different from saying (as the operationalist would) that there is no reality beyond measurements. The second thing a reality-of-states view should take seriously is *state reduction*. After a measurement on one of the two particles, standard quantum mechanics[4] no longer describes the atom-photon

[4]As is well known, there is no satisfying characterization of "measurement" within standard quantum

system by $|\Psi\rangle$, but by a separable state. A realistic view of the quantum state suggests that this change of description corresponds to a real physical change. This implies that the metaphysical significance of the sorting of atoms into subensembles depends on the temporal ordering of the measurements.[5] If the photon is measured prior to the atom's arrival at the screen, the subensembles correspond to real properties, because the photon measurement brings about the state reduction

$$|\Psi\rangle(x) \to |i\rangle\psi_i(x), \quad i \in \{1, 2, +, -\}, \tag{3}$$

so that each atom *actually is* in one of the ψ_i states prior to its hitting the screen. But if the time order of the two measurements is reversed, the atom *never* is in any one of these states, because (3) does not occur. Instead, the atom's arrival at the screen (at a location x_0) results in a state reduction of the form

$$|\Psi\rangle(x) \to [\alpha|1\rangle + \beta|2\rangle]\phi_{x_0}(x), \tag{4}$$

where $\phi_{x_0}(x)$ is a spatial wave function well localized at x_0.[6] In this case, assigning the atom to a subensemble depending on the result of the photon measurement implies nothing about the physical state of the atom, whether past or present.[7]

But then, isn't the appearance of definite (WW or DS) patterns within the subensembles somewhat miraculous? I do not think so. Given that the atom and the photon formed an entangled system up to the moment of the atom's detection, it is not so surprising that we can obtain interesting patterns by correlating the location of the atom's detection with the result of a posterior photon measurement. But this *correlating* is something the experimenter needs to *do*; the correlation is no longer "there", once the transition (4) has occurred.

4 Delayed-Choice Entanglement Swapping

We can now apply the foregoing considerations to the process of entanglement swapping, first proposed by Yurke and Stoler (1992). In the simplest case, this involves entangled

mechanics (Bell 1990). The reality-of-states view presented here is compatible with different solutions to this problem, e.g., the GRW theory or an Everett-type approach. In the latter case, state reduction is to be understood as a splitting of worlds due to environment-induced decoherence.

[5] This becomes problematic in relativistic settings, testifying to the unsolved problem of reconciling quantum non-locality with relativity (Maudlin 2011). I am here only interested in the non-relativistic case.

[6] For details about the coefficients α and β, see Englert et al. (1999, eq. 8). Of course, we could equally well express the photon state in the $|+\rangle$, $|-\rangle$ basis.

[7] Although I have derived this result within a view that takes state reductions as real events, it is interesting to note that the result is not peculiar to such a view. Hiley and Callaghan (2006b) analyze the situation from a Bohmian perspective, which does not regard state reductions as real, but instead assigns a definite trajectory to each atom. Their conclusion is analogous to mine, namely that the retrospective sorting into subensembles does *not* correspond to differences in the trajectories of the atoms.

pairs of two-state systems, which can be conveniently described by introducing the four so-called Bell states:

$$|\psi^{\pm}\rangle = \frac{1}{\sqrt{2}}[|0\rangle|1\rangle \pm |1\rangle|0\rangle], \quad |\phi^{\pm}\rangle = \frac{1}{\sqrt{2}}[|0\rangle|0\rangle \pm |1\rangle|1\rangle].$$

Now consider two independent sources, each one emitting a particle pair in the state $|\psi^{-}\rangle$. Denoting the two pairs by (A, B) and (C, D) respectively, the state of the complete system is given by

$$|\Psi\rangle = |\psi^{-}\rangle_{AB}|\psi^{-}\rangle_{CD}. \tag{5}$$

This is obviously a separable state, reflecting the fact that the two pairs are mutually independent. But now suppose that particles B and C are sent to the same location, where a Bell measurement is performed on them, such that their joint state is projected onto one of the four Bell states $|\psi^{\pm}\rangle_{BC}$, $|\phi^{\pm}\rangle_{BC}$.[8] To see how this affects particles A and D, we rewrite equation (5) by expressing $|\Psi\rangle$ in the basis given by the Bell states of the pairs (A, D) and (B, C):

$$|\Psi\rangle = \frac{1}{2}\left[|\psi^{+}\rangle_{AD}|\psi^{+}\rangle_{BC} - |\psi^{-}\rangle_{AD}|\psi^{-}\rangle_{BC} - |\phi^{+}\rangle_{AD}|\phi^{+}\rangle_{BC} + |\phi^{-}\rangle_{AD}|\phi^{-}\rangle_{BC}\right]. \tag{6}$$

Since the Bell states are orthogonal to each other, the Bell measurement on the (B, C) pair projects the state $|\Psi\rangle$ onto an entangled state of the (A, D) pair, for example:

$$\langle\psi^{+}|_{BC}|\Psi\rangle = \frac{1}{2}|\psi^{+}\rangle_{AD},$$

and analogously for the other Bell states. Thus particles A and D emerge as an entangled pair, although they never interacted with each other.

As Peres (2000) points out, this procedure can be carried out in a delayed-choice mode, such that the decision to perform a Bell measurement on the (B, C) pair may take place after any measurements on the (A, D) pair. But since particles A and D only become entangled with each other if the (B, C) measurement is actually performed, it seems that "entanglement is produced *a posteriori*, after the entangled particles have been measured and may no longer exist" (Peres 2000, 139). We have seen in the introduction that Healey takes this to undermine the idea that entanglement is a physical relation. To support his view, he offers the following *reductio ad absurdum*:

> To hold onto that idea in the context of this experiment would require one to maintain not only that *which* entanglement relation obtains between a pair of photons at some time, but also whether *any* such relation then obtains between them, depends on what happens to other independent systems later, after the pair has been absorbed into the environment. (Healey forthcoming, 24)

[8] This is an idealized description. In practice, a Bell measurement is unable to identify all of the four Bell states. For technical reasons, experiments usually focus on the singlet state $|\psi^{-}\rangle$ (Pan et al. 1998; Jennewein et al. 2002).

There is a clear parallel between this argument and the discussion in section 3, and this might seem to force me into accepting Healey's conclusion. In section 3, my unwillingness to accept retrocausation led me to reject the claim that the atom *really* went through either one of the slits (even if a WW measurement seems to tell us so). Should it then not also lead me to reject the claim that the (A, D) pair *really* either was or was not in an entangled state prior to the (B, C) measurement? But such an indefiniteness seems incompatible with the view that entanglement relations are real. This result would be particularly troubling in view of the fact that the notion of an entangled state played a crucial role in the account I defended in section 3.

Yet, a closer look reveals that the parallels between the two cases do not threaten realism about entanglement. Rather, they can be exploited to refute Healey's argument. In section 3, I showed that a delayed measurement of the photon results in a sorting of atoms into subensembles which do not correspond to any physical properties of the atoms. Precisely the same thing can happen in entanglement swapping: The Bell measurement on the (B, C) pair allows us to sort the (A, D) pairs into four subensembles corresponding to the four Bell states. Without delayed choice, this has physical significance, because each (A, D) pair *actually is* in such a state after the (B, C) measurement. But if the (A, D) measurements precede the (B, C) measurement, the (A, D) pair *never is in any of these states.* This is entirely compatible with the fact that evaluating the (A, D) measurements *within* a certain subensemble shows Bell-type correlations, just as the subensembles in section 3 showed interference or WW patterns.

Therefore, far from being committed to any indeterminism about entanglement (or any backward-in-time influences), a realistic view of the quantum state yields a perfectly clear assessment of what happens in entanglement swapping: If the (B, C) measurement occurs at a time the complete system is still in state $|\Psi\rangle$, it confers entanglement on the (A, D) pair, if it occurs at a later time, it does not.

5 Conclusion

I have argued that delayed-choice experiments do not undermine a realistic view of the quantum state. In the case of the double slit, we saw that they merely undermine a simplistic realism which unreflectively identifies the result of a WW experiment with a statement about which slit the particle went through. The quantum eraser and the case of delayed-choice entanglement swapping required a more careful treatment, because one needs to get clear about the metaphysical significance of the subensembles appearing in these experiments. Once this is achieved, a straightforward reality-of-states story can be told for these seemingly troubling cases.

This does, of course, not exclude non-realism about the quantum state. But it seems to me that the empirical success of quantum mechanics gives us at least some *prima facie* reason to view quantum states as describing an independent reality. The non-realist then needs an argument for the claim that this is a mistake. Wheeler, Scully et al., Mohrhoff and Healey all think that delayed-choice experiments furnish such an argument. This I have shown not to be the case.

The same dialectic applies, more specifically, to the metaphysics of entanglement. The various things physicists can *do* with entanglement support the intuition that there must be some reality to it. Against this, Healey argues that delayed-choice entanglement swapping implies an indeterminism about entanglement which is incompatible with realism. Having shown that the realist can avoid such an indeterminism, I do not see any reason to give up the realist intuition about entanglement.

References

Allori, V., S. Goldstein, R. Tumulka, and N. Zanghì (2008). On the common structure of Bohmian mechanics and the Ghirardi-Rimini-Weber theory. *British Journal for the Philosophy of Science 59*, 353–389.

Bell, J. S. (1990). Against 'measurement'. In A. I. Miller (Ed.), *Sixty-two years of uncertainty: historical, philosophical, and physical inquiries into the foundations of quantum mechanics*. New York: Plenum Press. Reprinted in J. S. Bell (2004): *Speakable and Unspeakable in Quantum Mechanics*. 2nd ed. Cambridge: Cambridge University Press, pp. 213-231.

Clifton, R. (2002). The subtleties of entanglement and its role in quantum information theory. *Philosophy of Science 69*, S150–S167.

Dürr, D., S. Goldstein, and N. Zanghì (1997). Bohmian mechanics and the meaning of the wave function. In R. Cohen, M. Horne, and J. Stachel (Eds.), *Experimental Metaphysics: Quantum Mechanical Studies for Abner Shimony*, Volume 193 of *Boston Studies in the Philosophy of Science*, pp. 25–38. Kluwer Academic Publishers.

Englert, B.-G., M. O. Scully, and H. Walther (1999). Quantum erasure in double-slit interferometers with which-way detectors. *American Journal of Physics 67*, 325–329.

Esfeld, M. (2004). Quantum entanglement and a metaphysics of relations. *Studies in History and Philosophy of Modern Physics 35*, 601–617.

French, S. (1989). Individuality, supervenience and Bell's theorem. *Philosophical Studies 55*, 1–22.

Healey, R. (forthcoming). Quantum theory: A pragmatist approach. *British Journal for the Philosophy of Science*. Preprint: http://philsci-archive.pitt.edu/8620/.

Hiley, B. J. and R. E. Callaghan (2006a). Delayed-choice experiments and the Bohm approach. *Physica Scripta 74*, 336–348.

Hiley, B. J. and R. E. Callaghan (2006b). What is erased in the quantum erasure? *Foundations of Physics 36*, 1869–1883.

Horodecki, R., P. Horodecki, M. Horodecki, and K. Horodecki (2009). Quantum entanglement. *Reviews of Modern Physics 81*, 865–942.

Jennewein, T., G. Weihs, J.-W. Pan, and A. Zeilinger (2002). Experimental nonlocality proof of quantum teleportation and entanglement swapping. *Physical Review Letters 88*, 017903.

Maudlin, T. (1998). Part and whole in quantum mechanics. In E. Castellani (Ed.), *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*, pp. 46–60. Princeton: Princeton University Press.

Maudlin, T. (2011). *Quantum Non-Locality and Relativity* (3rd ed.). Chichester: Wiley-Blackwell.

Mohrhoff, U. (1999). Objectivity, retrocausation, and the experiment of Englert, Scully, and Walther. *American Journal of Physics 67*, 330–335.

Pan, J.-W., D. Bouwmeester, H. Weinfurter, and A. Zeilinger (1998). Experimental entanglement swapping: Entangling photons that never interacted. *Physical Review Letters 80*, 3891–3894.

Peres, A. (2000). Delayed choice for entanglement swapping. *Journal of Modern Optics 47*, 139–143.

Scully, M. O. and K. Drühl (1982). Quantum eraser: A proposed photon correlation experiment concerning observation and "delayed choice" in quantum mechanics. *Physical Review A 25*, 2208–2213.

Scully, M. O., B.-G. Englert, and H. Walther (1991). Quantum optical tests of complementarity. *Nature 351*, 111–116.

Teller, P. (1986). Relational holism and quantum mechanics. *British Journal for the Philosophy of Science 37*, 71–81.

Timpson, C. G. (2010). Information, immaterialism, instrumentalism: Old and new in quantum information. In A. Bokulich and G. Jaeger (Eds.), *Philosophy of Quantum Information and Entanglement*, pp. 208–227. Cambridge: Cambridge University Press.

Wheeler, J. A. (1978). The 'past' and the 'delayed-choice' double-slit experiment. In A. R. Marlow (Ed.), *Mathematical Foundations of Quantum Theory*, pp. 9–48. New York: Academic Press.

Yurke, B. and D. Stoler (1992). Einstein-Podolsky-Rosen effects from independent particle sources. *Physical Review Letters 68*, 1251–1254.

Title: What is the “Paradox of Phase Transitions?”

Abstract:

I present a novel approach to the recent scholarly debate that has arisen with respect to the philosophical import one should infer from scientific accounts of “Phase Transitions,” by appealing to a distinction between “representation” understood as “denotation,” and “faithful representation” understood as a type of “guide to ontology.” It is argued that the entire debate of phase transitions is misguided for it stems from a pseudo-paradox that does not license the type of claims made by scholars, and that what is really interesting about phase transition is the manner by which they force us to rethink issues regarding scientific representation.

1. Introduction.

"Phase Transitions" (PT) include a wide variety of common and not so common phenomena in which the qualitative macroscopic properties of a system or a substance change abruptly. Such phenomena include, among others, water freezing into ice or boiling into air, iron magnetizing, graphite spontaneously converting into diamond and a semi-conductor transitioning into a superconductor. There exists a flourishing scholarly debate with respect to the philosophical import one should infer from the scientific accounts of phase transitions, in particular the accounts' appeal to the "thermodynamic limit" (TDL), and regarding how the nature of PT is best understood. It has become standard practice to quote the authoritative physicist, Leo P. Kadanoff, who is responsible for much of the advances in real-space Renormalization Group and in understanding PT, in order to better ground the puzzlement associated with PT:

> The existence of a phase transition requires an infinite system. No phase transitions occur in systems with a finite number of degrees of freedom. (Kadanoff 2000, 238)

If we add to the above that observations of boiling kettles confirm that finite systems do undergo PT, we conclude that rather odd paradox arises: PT do and do not occur in finite, and thus concrete and physical, systems. The above is taken as a basis for warranting such scholarly claims to the effect that PT are irreducible emergent phenomena (e.g. Lebowitz 1999, S346; Liu 1999, S92; Morrison 2012, 143; Prigogine 1997, 45), which necessitate the development of new physical theory (Callender 2001, 550), and inducing a wide array of literature that argues to the contrary (e.g. Bangu 2009; Batterman 2005, 2011; Butterfield 2011; Menon and Callender 2011; Norton 2011; Wayne 2009).

In this paper I would like to build on the works of Mainwood (2006) and Jones (2006) to further investigate what exactly is the "paradox" of PT, which is meant to license the type of scholarly conclusions and discussions noted above. It seems to me that a natural condition of adequacy for the particular claim that PT are emergent phenomena, as well as the more general debate that arises, is that there really be a bona fide paradox associated with PT. In other words, it really must be the case that a phase transition "is emergent precisely because it is a property of finite systems and yet only reducible to micro-properties of infinite systems," or more recently, that "the phenomenon of a phase transition, as described by classic thermodynamics cannot be derived unless one assumes that the system under study is infinite" (Lui 1999, S104; Bangu 2009, 488). Accordingly, in Section 2 I describe the paradox and suggest that much of the debate revolving around PT stems from it. In doing so, I appeal to Contessa's (2007, 52-55) distinction between "representation" understood as "denotation," and "faithful representation" understood as a type of "guide to ontology" (Sklar 2003, 427). Afterwards, I will continue to argue for a negative and a positive thesis. My negative thesis is that there really is no paradox of phase transitions and that in order to get a bona fide paradox, i.e. a contradiction, one must undertake substantial philosophical work and ground a type of "Indispensability Argument," akin to the kind appearing within the context of the Philosophy of Mathematics. Since none of the proponents of the PT debate undertake such work, and since indispensability arguments are highly controversial, I claim that the entirety of the debate, insofar as it is grounded in the paradox of PT, is utterly misguided and that the philosophical import that has been extracted from the case study of PT with regard to emergence, reduction, explanation, etc., is not warranted.

However, I also have a positive thesis. In Sub-section 2.1 I show how the paradox can be generalized and arises whenever a scientific account appeals to an "Essential Idealization"[1] (EI)—roughly, when a scientific account of some concrete physical phenomena appeals to an idealization in which, in principle, one cannot attain a more successful account of said phenomena by "de-idealizing" the idealization and producing a more realistic idealization. In doing so, I suggest in Section 3 that what is really interesting about phase transitions is the manner by which they illustrate the "Essential Idealization Problem," which is tightly connected to issues arising in the context of scientific representation and scientific realism. The upshot is that, insofar as proponents of the phase transition debate have been contributing to the EIP, certain aspects of the debate have been fruitful. Consequently, I outline various possible solutions to the EIP and the paradox of PT, which have been extracted from Butterfield (2011) and Norton (2011). I suggest that, although such solutions pave the road for further work to be done, it is questionable whether they are conclusive and exhaustive.

2. What is the "Paradox of Phase Transitions?"

In his 2001 paper, "Taking Thermodynamic Too Seriously," Craig Callender presents several allegedly true propositions that jointly induce a paradox concerning PT—that concrete systems can and cannot undergo PT:[2]

1. Concrete systems are composed of finite many particles N.
2. Concrete systems display PT.
3. PT occur when the partition function Z has a discontinuity.
4. The partition function Z of a system with finite many particles N can only display a discontinuity by appealing to the TDL.
5. A system in the TDL has infinitely many particles.[3]

Tenets 1-2 imply that concrete and *finite* systems display phase transitions while tenets 3-5 imply that only *infinite* systems can undergo a phase transitions. However, *contra* Bangu (2009), Callender (2001), Mainwood (2006), Jones (2006) and others, I contend that no contradiction arises by conjoining tenets 1-5. To see this, we must first distinguish between "concrete" phase transitions, on the one hand, and "abstract mathematical representations" of them, on the other hand.[4] To be clear, a "concrete system" would include a physical thermal system of type we find in the world or in a lab, while "abstract mathematical" just refers to pieces of math, e.g. a set with function defined on it. Also, I take the term "representation" here to be stipulated denotation

[1] Butterfield (2011) and Mainwood (2006) use the term "Indispensible," Jones (2006) uses "Ineliminable," and Batterman (2005, 2011) uses "Essential."

[2] The paradox of PT presented here in not the exact version presented in Callender (2001, 549). Instead, I present the paradox in a manner that is more relevant to my discussion. Several authors, such as Mainwood (2006, 223) and Jones (2006, 114-7), have undertaken a similar approach.

[3] For precise characterization of various forms of the TDL, see Norton (2011, sections 3 and 4) and reference therein.

[4] The distinction between concrete and abstract objects is a well-known. Abstract objects differ from concrete ones in the sense that they are non-spatiotemporal and causally inefficacious. Paradigm examples include mathematical objects and universals. Cf. Rosen (2001).

that is agreed upon by convention.[5] For instance, the notation "N" represents "the number of particles" (in a given system) in the sense that it *denotes* the number of particles. Second, notice that there are ambiguities with regards to whether the terms "PT" and "partition function" ("Z") in tenets 3 and 4 refer to concrete objects, or abstracts mathematical representations of them. As *concrete objects*, PT are concrete phenomena or processes that arise within concrete systems, while Z is some sort of concrete property of such systems. As *abstract mathematical representations*, both PT and Z are just pieces of mathematics that allegedly denote concrete objects. To avoid confusion, note that by "abstract PT" I only mean PT in the sense that an abstract Z displays a discontinuity. In the same manner, there is a clear ambiguity concerning the physical interpretation, i.e. the concreteness or abstractness, of the TDL. Thus, for example, if "PT" and "Z" in tenets 3 and 4 refer to abstract mathematical representations, as opposed to concrete objects, then there is no paradox: Concrete and finite systems display PT while abstract and finite ones do not. Just because abstract mathematical representations of concrete systems with finite N do not display PT, does not mean that concrete finite systems do not display PT. Alternatively, if "PT" in tenets 3 and 4 do refer to concrete PT, it also does not immediately follow that there is a paradox. Rather, what follows is that concrete PT "occur" when abstract representations of them display various abstract properties, such as a discontinuity in Z and an appeal to the TDL. One might wonder what explains this particular correlation between discontinuities in abstract representational partition function and concrete phase transitions. However, *prima facie*, there is no paradox.

The point is that without adding additional tenets that make a claim about the relation between, one the one hand, concrete PT occurring in physical systems and, on the other hand, the abstract mathematical representation of concrete PT, which arise in scientific accounts of PT, no paradox arises. In the following sub-section I will add such additional tenets in hope to further shed light on the central philosophical issue that arises in the context of PT. To end, it is worth noting that, if my claim about there being no paradox is sound, then the entire the debate revolving around PT, insofar as it is grounded in the paradox of PT as it is stated above, is unmotivated and utterly misguided. In particular, notice that the various positions expressed with regards to the debate can be delineated by identifying which tenet of the paradox a particular proponent denies or embraces. Authors such as Lebowitz (1999, S346), Liu (1999, S92), Morrison (2012, 143) and Prigogine (1997, 45) can be read as embracing tenet 3 and identifying PT as a kind of non-reductive emergent phenomena. Contrasting attitudes have been voiced by Wayne (2009), where Callender (2001) and Menon and Callender (2011) explicitly deny that phase transitions are irreducible and emergent phenomena by rejecting tenet 3. Butterfield (2011) can be read as both denying and embracing tenet 3, in an effort to reconcile reduction and emergence. Norton (2011) can be understood as denying tenet 5. I refer the reader to Mainwood (2006, 223-237), who presents an exposition of this type of delineation—i.e. a classification of scholarly attitudes to the nature of phase transition grounded in the paradox. For my purposes what is important is to identify that *the large majority, if not all, of the phase transition debate stems from the phase transition paradox.*

2.1 The bona fide Paradox of Phase Transitions and its Generalization

[5] Cf. Contessa (2007, 52-55) and references therein.

The key ingredient necessary to engender a bona fide paradox is for a particular kind of correspondence relation to hold between abstract representations and concrete systems. To make this point clear we must appeal to a further distinction. While I take "representation" to be stipulated denotation, by "faithful representation" I mean a representation that allows agents to perform sound inferences from the representational vehicle to the target of representation (Contessa 2007, 52-55). That is to say, a faithful representation allows agents to make inferences about the nature of the target of representation. Thus, it acts as a kind of "guide to ontology"[6] since it accurately describes aspects of the target of representation. In other words, a faithful representation is one in which the vehicle and target of representation resemble each other in some manner, e.g. they share some of the same, or approximately same, properties and/or relations. The classic example here is a city-map, which is a faithful representation of a city because it allows us to perform sounds inferences from the vehicle to the target, i.e. from the map to the city. This is so because both the vehicle and the target share various properties. For instance, if two streets intersect in the map, then they also intersect in the city. That is to say, intersecting streets in the map correspond to intersecting streets in the city. Therefore, the map acts as a type of ontological guide accurately describing the city, e.g. there *really* are intersecting streets in the city. It is worth noting that my account potentially differs from Contessa (2007), who isn't clear about the ontological aspect of faithful representations. Contessa (2007) differentiates from "epistemic representation," from which *valid* inferences can be drawn, and faithful ones that permit sound inferences. Whether or not such inferences come with ontological baggage depends on whether they are about the target itself. On my account, faithful representations license sound inferences about the target itself and hence they the fix the ontology of the target.

With this distinction in hand, if we add a tenet that says the abstract representational discontinuities representing phase transitions are *faithful* and hence correspond to concrete physical discontinuities we do get a genuine contradiction. This is so because if systems are composed of finite many particles, which is the case within the context of the atomistic theory of matter conveyed in tenet 2, then it makes no sense to talk of concrete discontinuities. The notion of concrete discontinuities presupposes that matter is a continuum so that there can be an actual discontinuity. Otherwise, an apparent discontinuity is actually the rapid coming apart of particles and not a real discontinuity. Consequently, adding a tenet as the one just described amounts to claiming that systems are not composed of finite many particles and so we get: Concrete systems are and are not composed of finite many particles N.

In a similar manner, one can engender a kind of paradox by reifying the TDL through an appropriate correspondence relation. For instance, one could add the tenet that an appeal to the TDL, which could be interpreted as a type of continuum limit faithfully representing an *abstract* system, in fact faithfully represents a *concrete* system. Thus, we deduce the claim that concrete systems are and are not composed of finite many particles N (in the sense that the ontology of concrete systems is both atomistic and that of a continuum, i.e. not atomistic).

The source of the problem of PT seems to be that the mathematical structure that scientifically represents concrete PT—a discontinuity in the partition function—is an artifact of an idealization (or an approximation)—the TDL—which is essential in the sense that when one

[6] Cf. Sklar (2003, 425).

"de-idealizes" said idealization, the mathematical structure representing PT no longer exist.[7] Accordingly, I would like to suggest that what is really interesting about PT is the manner by they might shed light on the nature of scientific representation and idealization. In particular, notice that once concerns regarding representations are incorporated, the paradox of PT can be generalized by making use of the concept of an EI:

1. Concrete systems include a concrete attribute A.
2. Concrete systems display a concrete phenomenon P.
3. P is scientifically-mathematically represented by P'.
4. P' can only arise by appealing to an idealizing limit I.
5. A system in the idealizing limit I includes an attribute $A^{\approx}$ such that $A \neq A^{\approx}$.
6. P' faithfully represents P.

Tenet 1 and 2 imply that concrete systems are A and display P. Tenets 3-5 imply that P is scientifically represented by P', which presupposes $A^{\approx}$. Tenet 4 encompasses our EI since any de-idealization of I will render P' nonexistent. So far there is no contradiction. But, when one adds the correspondence relation described by tenet 6, a bona fide paradox arises: Concrete systems are and are not A (since they are A and they are $A^{\approx}$and $A \neq A^{\approx}$). What is important to notice is that tenets 1 and 2 are claims about *concrete* systems, wherein tenet 2 identifies the concrete phenomenon to be scientifically accounted for, while tenets 3-5 are claims about *abstract* scientific accounts of concrete systems, and it is tenet 6 that connects the abstract with the concrete via faithful representation, thereby engendering a genuine paradox. The question, of course, is why would one endorse tenet 6? The answer is that without tenet 6 the entire scientific account of the concrete phenomenon in question seems somewhat mysterious to anyone with non-instrumental sympathies. In particular, those with realist intuitions will want to unveil the mystery with a correspondence relation that tells us that our abstract scientific accounts gets something right about the concrete world. But how would one argue for a correspondence relation along the lines of 6? It seems to me that, given the "essentialness" aspect of the idealizing limit that arises in tenets 3 and 4, the only way to justify tenet 6 is by appeal to an indispensability argument.[8] In other words, something of the sort:

1) A scientific account of some concrete phenomena appeals to an idealization(s) and refers to idealized abstract objects.
2) The idealization appealed to is essential to the scientific account in the sense that any de-idealization renders the scientific account less successful and the idealized abstract object nonexistent.
3) Hence, the idealization appealed to, and the idealized abstract objects made use of, are *indispensible* to the account.
4) Thus, as scientific realists, we ought to believe that such abstract idealized objects do exist and are concrete. Further, the ontological import of such idealizations is true of concrete systems, on pain of holding a double standard.

[7] For a more precise statement to this effect see Butterfield's (2011, 1123-1130) and Mainwood's (2006, 216-218) discussion of Lee-Yang Theory and KMS states.
[8] For a survey of the Indispensability Argument of mathematics and a defense see Colyvan (2001).

Said differently, and in the specific cases of PT, since reference to a discontinuity in Z is indispensible to scientific accounts of PT, and since these discontinuities only arise by appealing to EI, we ought to believe in the existence of concrete discontinuities.

Thus, in contrast to many of the scholars engaged in the phase transition debate, which assume that there is a paradox and then continue to attempt to dissolve it by some manner or other, I claim that in order to get a genuine paradox one needs to justify a correspondence relation (such as the one appearing in tenet 6) by appealing to an indispensability-type argument. Since cogent indispensability-type arguments require serious philosophical work and are very much controversial, and since no author engaged in the phase transition debate has undertaken such work, it follows that much of the controversy revolving around phase transitions is not well-motivated. That is to say, claims to the effect (i) that PT are or are not emergent, (ii) that they are or are not reducible to Statistical Mechanics (SM), and (iii) that they do or do not refute the atomic theory of matter, are grounded in a frail foundation that does not licensed such significant conclusions.

One might worry that, contrary to my claims, a bona fide paradox of PT can arise on the epistemological level by conceding to a set of tenets from which it is possible to deduce that SM does and does not govern phase transitions. The idea here is to argue that "SM-proper" is not licensed to appeal to the TDL and so SM-proper does not govern PT. However, the objection continues, it is generally assumed that SM is the fundamental theory that governs PT. Thus, we have a paradox and the natural manner by which to dissolve it is to argue that SM-proper does indeed have the tools to account for PT (Callender 2001, Menon and Callender 2010), or else to claim that PT are emergent. In reply, it is far from clear to me that SM-proper is not licensed to appeal to the TDL, and so that it does not govern PT. In fact, there are reasons to think that the TDL is 'part and parcel' of SM-proper because (a) it is common practice to appeal to the TDL in modern approaches to SM, and (b) the TDL is used in SM not only to account for phase transitions but to account for, among others, the equivalence of SM ensembles, the extensivity of extensive thermodynamic parameters, Bose condensation, etc. (Styer 2004). In addition, (c) all the best scientifically accounts of PT, and these include mean field theories, Landau's approach, Yang-Lee theory and Renormalization Group methods, represents PT as discontinuities by appealing to the TDL, and (d) the large majority of empirically confirmed predictions of SM, within the context of PT and beyond, appeal to the TDL.

Moreover, even if it was the case the SM-proper is not licensed to appeal to the TDL, no contradiction would arise. Rather, it would just be a brute fact that SM-proper does not govern phase transitions and "SM-with-the-TDL" does. If then it is claimed that the ontologies of SM-proper and SM-with-the-TDL are radically different so that indeed there is a paradox, we must notice that such a claim amounts to no more than reviving the paradox at the level of ontology, and hence my discussion in this section bears negatively on this claim.

Last, the claim that PT are emergent because SM-proper cannot account for them seems to replace one problem—PT are not governed by the fundamental theory—with another problem—PT are emergent. How does dubbing PT "emergent" illuminate our understanding of them or of their scientific accounts? How is this philosophically insightful? Accordingly, I endorse Butterfield's (2011) description of emergence as novel and robest mathematical structure that arises at a particular limit, as opposed to a failure of intertheoretic reduction of some sort. It is worthwhile to note that the insistence on the indispensibility of taking such limits

for the purpose of emergence understood in this manner has been repeatedly stressed by, e.g., Batterman (2005, 2010, 2011).

3. The Essential Idealization Problem.

The above discussion points to what I consider to be the central philosophical issues arising out of the debate concerning PT. First, the discussion regarding (i) the need for a correspondence relation between our abstract scientific-mathematical representations and concrete systems, (ii) the appeal to the concept of "faithful representation," and (iii) the identification that the phase transition paradox can be generalized to any scientific account that appeals to EI, demonstrates that a solution to the following problem is needed:

> The Essential Idealization Problem (EIP) — We need an account of how our abstract and essentially idealized scientific representations correspond to the concrete systems observed in the world and we need a justification for appealing to EI's, i.e. an explanation of why and which EI's are successful, which does not constitute a de-idealization scheme.[9]

To this effect Batterman (2005, 2010, 2011) has made progress by explaining that it is not at all clear that traditional mapping accounts of scientific and mathematical representation work in cases of EI. In particular, this is so because the abstract mathematical structure doing the representational work does not "latch on," and so is not partially isomorphic or homomorphic, to any concrete physical structures in the external word. Moreover, insofar as the physical world constrains scientific representations, there are reasons to think that consideration of scale size, in which the phenomenon of concern occurs, plays an important role in modeling and scientifically representing such phenomenon.

Second, the discussion of indispensability makes it clear that the mystery revolving around the EIP is truly mysterious for those with scientific realist sympathies and, in fact, may threaten certain conceptions of realism. This follows because, insofar as arguments like the "no miracles argument" and "inference to best explanation" are cogent and give us good reason to believe the assertions of our best scientific accounts, including those about fundamental laws and unobservable entities, then in the case of accounts appealing to EI, these arguments can be used via an Indispensability Argument to reduce the realist position to absurdity. What is needed is a realist solution to the EIP and thus a realist account of PT.

In fact, such potential solutions to paradox of PT can extracted from two recent contributions to the debate: Butterfield (2011) and Norton (2011). Although it is beyond the scope of this paper to treat these contributions thoroughly, I will end by discussing them shortly in effort to support my suggestion that, although such solutions pave the road for further work to be done, it is questionable whether they are conclusive and exhaustive.

Butterfield (2011) grants that the TDL is "epistemically indispensable" for the emergence of the novel and robust mathematical structure that is used to represent PT, but denies that any paradox emerges because the limit is not "physically real." Using the terminology expressed

[9] Mainwood (2006, 214-5) also identifies a similar problem but in a context that is different from mine, and his solution (238), endorsed by Butterfield (2011), misses the central issue discussed here.

here, the discontinuities in Z play a representational role but not a *faithfully* representational one. The question arises, how come unfaithful representations work so well? To that end, Butterfield (2011, Section 3) appeals the distinction, also used by Norton (2011, Section 3), between "limit quantities" or "limit properties," i.e. the limits of properties, and "limit system," i.e. the system at the limit. He continues to argue that the behavior of certain observable properties of concrete finite systems, e.g. magnetization of a ferromagnet, smoothly approaches the behavior of the corresponding properties of abstract infinite systems. Moreover, it is the large N behavior, not the infinite N, which is physically real.

Norton (2011) suggest that by viewing the TDL as an "approximation"—an inexact description of a target system, instead of an "idealization"—a novel system whose properties provide inexact descriptions of a target system, we can diffuse any problems that might arise. Within the context of our discussion, Norton's idea is that no paradox can arise if the TDL is an approximation since approximations do not refer to novel systems whose ontology might be drastically different from the target systems, thereby engendering a paradox once we add an appropriate correspondence relation. In a similar manner to Butterfield (2011), his justification for appealing to such an approximation is pragmatic: the behavior of the non-analytic Z belonging to an infinite system, is approached by an analytic Z corresponding to finite system with large N.

From my viewpoint, this cannot be the whole story. First, both accounts seem to ignore that it is a mathematical structure that arises only in the limiting system that is doing the representational work for us. Moreover, the accounts seem to suggest that we must revise our definition of PT as occurring when the partition function has a discontinuity, and substitute it with something along the lines of "PT occurs when various thermodynamic potentials portray sufficiently extreme gradients." The weakness of this suggestion is that we have substituted a precise characterization of PT, with a vague one. But more problematic is the idea that we should be able to construct a finite N system that has a, say, Helmholtz free energy with an extreme gradient, which does evolve into a discontinuity once the TDL is taken.[10] Second, the Butterfield-Norton approach outlined above seems incomplete for it does not give us an account of why it is that the concrete external world constrains us to model and scientifically represent certain phenomena with mathematical structures that only emerge in limiting systems whose ontology does not correspond to that of the fundamental theory. For this purpose, talk of "mathematical convenience," "empirical adequacy," and "approximation" (understood as a purely formal procedure) misses what seems to be the truly intriguing features of PT. My suggestion is that we can further advance our understanding of PT, and similar phenomena that gives rise to the EIP, by attempting to amend accounts like Butterfield's (2011) and Norton's (2011) with some of the key insights of Batterman (2005, 2011) regarding what mathematical techniques one must appeal to in order to properly represent certain kinds of phenomena.

[10] Mainwood (2006, 232) makes the same point.

References

Bangu, S. (2009) "Understanding Thermodynamic Singularities: Phase Transitions, Date and Phenomena." *Philosophy of Science* 76:488-505.

Batterman, R. (2005) "Critical phenomena and breaking drops: Infinite idealizations in physics." *Studies in History and Philosophy of Modern Physics* 36:225-244

———. (2010) "On the Explanatory Role of Mathematics in Empirical Science." *British Journal for the Science* 61(1):1-25.

———. (2011) "The Tyranny of Scales." http://philsci-archive.pitt.edu/8678/.

Butterfield, J. (2011) "Less is Different: Emergence and Reduction Reconciled." *Foundations of Physics* 41(6):1065-1135.

Contessa, G. (2007) "Scientific Representation, Interpretation and Surrogative Reasoning." *Philosophy of Science* 74:48-68.

Colyvan, M. (2001) *The Indispensability of Mathematics*. NY: Oxford University Press, Inc.

Kadanoff, L. P. (2000) *Statistical Physics: Statics, Dynamics and Renormalization*. Singapore: World Scientific.

Lebowitz, J. L. (1999) "Statistical Mechanics: A Selective Review of Two Central Issues." *Reviews of Modern Physics* 71(2):S346-S357.

Jones, N. J. (2006) "Ineliminable Idealizations, Phase Transitions and Irreversibility." PhD diss., Ohio State University.

Lui, C. (1999) "Explaining the emergence of cooperative phenomena." *Philosophy of Science* 66 (Proceedings): S92–S106.

Menon, T. and C. Callender. (2011) "Turn and Face the Strange… Ch-ch-changes: Philosophical Questions Raised by Phase Transitions." http://philsci-archive.pitt.edu/8757/.

Morrison, M. (2012) "Emergent Physics and Micro-Ontology" *Philosophy of Science* 79:141-166.

Mainwood, P. R. (2006) "Is More Different? Emergent Properties in Physics." PhD diss., Oxford University.

Norton, J. D. (2011) "Approximations and Idealizations: Why the Difference Matters." http://philsci-archive.pitt.edu/8622/.

Prigogine, I. (1997) *The End of Certainty*. The Free Press, New York.

Rosen, G. (2001) "Abstract Objects." In *Stanford Encyclopedia of Philosophy*, ed. Edward N Zalta. Stanford, CA: Stanford University, http://plato.stanford.edu/entries/abstract-objects/.

Sklar, L. M. (2003) “Dappled Theories in a Uniform World.” *Philosophy of Science* 70:424-441.

Styer, D. F. (2004) “What Good is the Thermodynamic Limit?” *American Journal of Physics* 72(1):25-29.

Wayne, A. (2009) “Emergence and Singular Limits.” *Synthese* 184(3):341-356.

Abstract and Complete

Abstract

There are two notions of abstraction that are often confused. The material view implies that the products of abstraction are not concrete. It is vulnerable to the criticism that abstracting introduces misrepresentations to the system, hence abstraction is indistinguishable from idealization. The omission view fares better against this criticism because it does not entail that abstract objects are non-physical and because it asserts that the way scientists abstract is different to the way they idealize. Moreover, the omission view better captures the way that abstraction is used in many parts of science. Disentangling the two notions is an important prerequisite for determining how to evaluate the use abstraction in science.

I. Introduction

The west pediment of the Parthenon is a physical object that exists in space and time, but it is also triangular. We say that the west pediment is concrete, but that triangles are abstract. What accounts for this difference? The received view in philosophy of science is that an object is abstract when it is not concrete (e.g. Cartwright 1994). Call this the *material view* of abstraction. The problem with the material view is that it implies that abstract objects are not physical. However, scientists often work with systems that are abstract but also physically instantiated. For example, experiments conducted in greenhouses abstract away from properties such as the color of the plants in question and whether or not they are subject to herbivory. Nonetheless, the plants in these experiments are concrete particulars like the west pediment of the Parthenon and unlike triangles. Moreover, the material view blurs the distinction between abstraction and idealization, as idealized objects are not concrete. For example, assuming that a population is infinite is common practice in models of population genetics, yet no actual population in the world is infinite. In this sense, infinite populations are like triangles

and unlike the west pediment of the Parthenon. The problem is that the main goal of proponents of the material view is to defend abstraction from critics who argue that both abstraction and idealization involve distortion, hence they are not distinct processes (e.g. Humphreys 1995). Unfortunately, the material view of abstraction undermines the force of their arguments against the critics.

Thomson-Jones defends a different view of abstraction where abstraction means the omission of irrelevant parts and properties from an object or system (Jones 2005).[1] I will call this the *omission view*. Here, abstraction and idealization are distinct because idealization requires the assertion of a falsehood, while abstraction involves the omission of a truth (ibid). Thus, while both idealization and abstraction can result in the distortion of a system, the distortion is very different in each case. When we abstract, we do not describe the system in its entirety, so we are not telling the whole truth. However, when we idealize, we add properties to the system that it does not normally possess. Therefore, our description of an idealized system contains falsehoods.

Both the material and omission views about abstraction are relevant to parts of scientific inquiry, but it is important to keep them distinct. If we fail to do so and lump abstraction together with idealization, we are in danger of trivializing an important aspect of science. I will argue that the notion of abstraction that is relevant to models, modeling, experiments, and target system construction (Godfrey-Smith 2006) is a version of the omission view. Specifically, this is the view that abstraction is the opposite of completeness. We start off with a complete object or system, one that has all its parts

1 Cartwright also defends this view in places, yet she uses the two notions interchangeably (Cartwright 1994). This implies that she views the material and omission views as two different aspects of the same notion instead of two distinct notions of abstraction.

and properties. When we abstract, we omit the parts and properties that are irrelevant for our purposes. An important implication of this view is that the outcomes of the process of abstraction can be concrete and physical.

II. The use of Abstraction in Science

The material view of abstraction is intuitive and deeply entrenched. Prime examples of abstract objects are mathematical objects such as numbers and triangles, which are not physically instantiated. Examples of abstract objects in other disciplines are concepts and ideas which are not tangible (e.g., fairness, evil, superego). Interestingly, in many of these cases, we can arrive at these objects through the process of omission. For example, we can start off with two roses, omit properties such as color, smell, photosynthetic capacity, chemical composition and so on, until we arrive at the number two. Historically, philosophers writing on abstraction (e.g. Aristotle and Locke) have held versions of the material view but explained how we arrive at abstract objects with the omission view (Rosen 2009, Cartwright 1994). It is not surprising, therefore, that the two views of abstraction are often lumped together as aspects of the same notion.

However, the use of abstraction in science is often quite different. Scientists often omit a number of parts and properties from a system, yet do not treat the resulting systems as immaterial or intangible. In the remainder of this section I will give some examples systems used by scientists that are both abstract and concrete. The first is an experiment from plant ecology, aimed at determining the cause of competition between two plants. In this experiment, Jarchow and Cook (2009) conducted a series of

Abstract and Complete

experiments with the invasive aquatic cattail species *Typha angustifolia* and the native wetland species *Bolboschoenus fluviatilis*, which inhabit North American lakes. They took specimens from both species back to the greenhouse and grew them in a single controlled environment. The results showed that *T. angustifolia* had a competitive advantage over *B. fluviatilis* because of allelopathy (the exudation of toxins from its roots). These toxins inhibit the growth of the native species (with a resulting 50% reduction in biomass) which allows the invader to soak up the limited nutrients in the soil. Above ground, the invader rapidly increases in size and shades the native species, which further reduces its growth rate.

It seems strange to think of this experiment as an abstract system, if we retain the idea that abstract objects are immaterial. The system of the plants in the greenhouse is as tangible and physically instantiated as the plants in the lake ecosystem. However, by bringing the plants into the greenhouse, the scientists are excluding all the other parts and properties of the lake ecosystem. The experiment, conducted in a simplified environment, allowed the scientists to identify the existence of competition between the two plants and to isolate the cause of the competitive advantage of *T. angustifolia.* They achieved this by being able to isolate the important factors from the system and omitting or parametrizing the other, irrelevant factors. In other words, the scientists started off considering a complete system with all its parts and properties (the lake ecosystem) and ended up with a system with fewer parts (fewer individuals from fewer species) and properties (the particular plants are not thought of as prey, or as contributing to the uptake of atmospheric CO_2).

Abstract and Complete

Moreover, this example is not a one-off case. The very nature of experimentation in ecology is based on the idea that ecosystems are very complex and identifying the most important causal factors that lead to ecological phenomena involves controlling and parametrizing other factors. The same is true of experiments in evolutionary biology. Geneticists test mutation rates in populations of *E. coli* and *Drosophila* in controlled laboratory settings. The point of those experiments is to isolate the genetic factors that affect mutation rates, without the compounding or mitigating effects of developmental and environmental variation. Even further afield, experiments in psychology are conducted in controlled environments, with the aim of minimizing irrelevant effects.

Abstraction is also an important step in modeling. As with experimentation, when scientists model a particular phenomenon in a system, they do not model the entire system but a subset of parts and properties of that system. The identification of which parts of the system are important and the omission of those parts that are not, is another example of the process of abstraction.

I will illustrate with an example from population ecology. The marmots of Vancouver Island (*Marmota vancouverensis*) are classified as critically endangered. It is estimated that their population has dropped 80%-90% since the 1980's and currently consists of roughly 200 individuals (Brashares et al. 2010). Ecologists studying these social rodents wish to understand how to bring back the population from the brink of extinction. In order to that, they must understand the causes of the decline in the marmot population. A good place to start is to look at a standard model of population

Abstract and Complete

growth and check if the actual marmot population deviates from the model (this was the exact strategy undertaken by Brashares and colleagues) (ibid). There are a number of models in ecology which measure population growth; the logistic growth model (originally developed in 1838 by Pierre Verhulst) is often used in the early stages of a study, because it is not entirely unrealistic (as it takes into account the effect of density on population growth) but at the same time it is quite simple (Fig 1).

Fig 1. Logistic Growth Model

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \qquad (1)$$

(N) is the number of organisms in population. (r) is the intrinsic growth rate of the population. (K) is the carrying capacity of the environment: the total number of organisms a particular environment can support.

This model measures how the growth rate of a population (*N*) is limited by the density of the population itself. (*r*) is the intrinsic growth rate, the maximum possible growth rate of the population. It is roughly equivalent to the number of deaths in the population subtracted from the number of births in that population.[2] The second important component of the model is (*K*), the *carrying capacity* of the environment. (*K*) imposes the upper limit on population growth because it is the maximum number of

2 Different species have different intrinsic growth rates; for example, large mammals such as elephants reproduce slowly and therefore have a low (*r*) whereas most insects and plants have high reproductive rates and therefore have a high value of (*r*).

Abstract and Complete

individual organisms that a particular environment can support. Factors that affect (*K*) are the environment's resources, yet they vary across environments and species.[3]

There are two sets of abstractions from the Vancouver Island (VI) ecosystem that need to occur so that the population growth of actual marmots can be compared with the prediction of the logistic growth model. The first is the elimination of *parts* that are not relevant. This includes the elimination of all units that are not relevant for measuring the population growth of marmots. The other animals, most of the plants on VI, and inanimate parts such as the marmot burrows will be omitted. The only other parts of the system that will be included are the plants that the marmots feed on (for example, cow parsnips, Kinnikinnick-fruit and huckleberries). The second set of abstractions concerns the *properties* that are relevant for the experiment or model. Properties such as eye color, fur length and fur color will not be relevant, because they do not affect short-term population growth. On the other hand, properties such as sex, time spent foraging and metabolic rate are relevant because they determine (*r*) the intrinsic growth rate of the marmot population.

With these abstractions in place, scientists were able to figure out that the growth rate of the marmot population on VI was falling, despite being far from close to the carrying capacity of the island. The reason for this is a phenomenon known as the Alee effect (named after Warder Clyde Allee who first described it). This effect occurs in

3 For example, in the case of plants, access to sunlight is very important, as are elements such as phosphorus and nitrogen. The amount and availability of each of these factors in the system will affect the (*K*) of plant populations. For many social mammals, space is very important as it affects the location of territory or the number of nesting sites. For example, the size of beaver populations in an area is partly determined by where each family can build its dam (and each dam's proximity to other dams).

small populations when a fall in population density decreases the growth rate instead of increasing it. Brashares et al. found that this instance of the Allee effect was caused by a 'social meltdown' (ibid). Unlike other marmots, VI marmots are very social and the decline in population leads to difficulty in finding mates, which reduces the growth rate even more.

This example is aimed at showing that abstraction is an integral part of modeling in science. In the paper, the logistic growth model is compared with the actual population of marmots, considered in isolation from the other parts of the ecosystem (ibid). There is no reason to think that the collection of marmots and the properties of their population is not concrete. Nonetheless, the population of VI marmots has fewer parts than the entire ecosystem on VI. In this second sense, it is more abstract that the entire VI ecosystem.

To recap the argument so far, there are two views of abstraction: the material view and the omission view. On the material view abstract objects are immaterial. On the omission view abstract objects are simply incomplete, and can be either material or immaterial. The two views are easily confounded because immaterial abstract objects result from the process of omission. However, there are a number of examples in science where the process of omission leads to physical objects or systems. Thus, the material view cannot account for all the objects or systems that arise from the process of omission. In contrast, the omission view accounts for all systems that result from omission, irrespective of whether or not they are concrete. Thus, if we want a single, unified notion of scientific abstraction, then we should opt for the omission view.

Abstract and Complete

III. Abstraction and Idealization

In the introduction, I mentioned another criticism of the material view of abstraction, namely that abstraction and idealization are not distinct concepts and they can be used interchangeably to signify any distortion in the scientific representation of a phenomenon. This view, endorsed explicitly by some (Humphreys 1995) and implicitly by many more (McMullin 1985), implies that there is no real or interesting distinction between abstraction and idealization. The two processes are thought to be inextricably linked, if not identical, and attempting to separate them results in confusion. The main proponent of the material view of abstraction is Paul Humphreys, who argues that in order to talk about abstract systems we usually have to represent them in some manner, and this representation will not be concrete (Humphreys 1995). However, idealized systems are also representations that are not concrete. According to Humphreys, the two types of representations are, therefore, not easily distinguishable.

This diagnosis is quite apt. Cartwright (the main proponent of the material view) states that when we idealize, we start off with a concrete object and "mentally rearrange some of its inconvenient features -some of its specific properties- before we try to write down a law for it" (Cartwright 1994 187). In contrast, when we abstract, we strip away properties from a system "in our minds" (Cartwright 1994). Thus, for example, when we omit all the irrelevant properties from the west pediment of the Parthenon, we are left with the shape of a triangle. This shape cannot be a true triangle though, as it is not a perfect geometrical shape. This is because the west pediment contains imperfections which are retained in the process of abstraction. According to Cartwright, this does not

really matter, as we can pretend that the abstract shape is a true triangle. The imperfections are already present in the real system and are not the result of our abstraction. In addition, these imperfections are themselves insignificant, and for all intents and purposes the abstract triangle is close enough to a true triangle. Thus, despite the imperfections retained in the process of abstraction, we are close enough to the real systems that we are entitled to pretend that our abstract shape is a true triangle.

The problem, as Humphreys points out, is that once we start pretending what a system is like, we blur the lines between abstraction and idealization. We cannot legitimately focus on the triangle's geometrical properties because an imperfect concrete triangle will remain imperfect after we abstract. If we want our abstract triangle to have geometric properties, then we have to *add* them to abstract triangle. In the case of *true* abstraction all the properties of the abstract object already exist in the real world. Hence, as soon as we start pretending, we are adding properties to our system that the material triangle does not have. In other words we are misrepresenting, or distorting the system. If this is the case, then abstraction and idealization seem very similar. To put the point differently, adding geometrical properties to a triangle is very much like assuming that a population in biology is infinite. No triangle in the actual world is perfect, just as no population of organisms in the world is infinite. In both cases, misrepresenting the system by adding properties is extremely useful, as it helps us model the system with the use of mathematics. Nonetheless, misrepresentation of a system, *according to proponents of the material view*, counts as idealization.

Abstract and Complete

I agree with Humphreys that this is an important problem for the material view of abstraction. As soon as we disassociate abstract objects from concrete objects, then we are abstracting 'in our minds' and representing them imperfectly. However, this criticism loses its force when pitted against the omission view of abstraction. On this view, abstraction is 'mere omission', i.e., we only abstract properties that are irrelevant for our system (Jones 2005). In the case of the west pediment, these properties are the pediments color, the fact that it contains statues, that is made of marble. What we are left with is a concrete shape that is also triangular. Importantly, this triangular shape is *not* a true triangle, it is simply approximates a true triangle. Mere omission cannot give rise to an immaterial true triangle from the imperfect and concrete pediment.

On the omission view, abstracting from the west pediment is like abstracting parts and properties from the VI ecosystem in order to explain the population size of the VI marmots. In the case of VI, the ideal population is represented by the model which is compared to the size of the actual population of marmots. Similarly, a true triangle can be compared to the actual approximately triangular shape of the west pediment. The difference between the material and omission views is that in the latter, there is no pretending. On the omission view, we can identify differences are between the abstract and ideal systems. Hence abstraction and idealization can be kept distinct.

A distinct criticism which does bear against the omission view attempts to assimilate abstraction to idealization because both fundamentally involve distortion.[4] The idea is that omitting aspects of a system results in the misrepresentation of the

4 This criticism stems from the view that idealization is not a unified, singular concept. Proponents of this view (Weisberg 2007, Frigg & Hartmann 2009) believe instead that there are different kinds of idealization in science and that abstraction is subsumed under one of these kinds of idealization.

system. Consequently, abstraction is a special case of idealization. In other words, no parts or properties of a system are strictly speaking 'irrelevant', hence they cannot be omitted from without the system being distorted. Omission necessarily results in distortion, because systems in nature are irreducibly complex. For example, ecosystem ecology is subfield of ecology that advocates holistic approach that views ecosystems as wholes or even individuals (Odenbaugh 2007). This is in direct contrast to the subfield of population ecology, where population dynamics are thought to capture and explain ecological phenomena. The big difference between the two approaches is that population ecologists work with more abstract models, as they omit a number of factors (especially abiotic factors) as irrelevant. On the other hand, ecosystem ecologists think that omitting abiotic factors from complex ecosystems results in overly simplistic models. The problem with that is that various processes which involve abiotic factors are themselves omitted or misrepresented, which in turn gives a distorted view of the way an ecosystem functions. In other words, it is the omission of factors from the system that leads to its misinterpretation.

Thomson-Jones attempts to avoid this problem by restricting abstraction to precisely those omissions that do not result in misrepresentation (Jones 2005). As stated above, a 'mere omission' does not misrepresent a particular feature of a system because it retains 'complete silence' with respect to whether the system contains the feature (ibid). So if an omission results in a misrepresentation, then it is not the type of omission that is part of abstraction. The problem is that the criticism presented here is much stronger. The criticism denies the possibility of 'mere omission' altogether.

Abstract and Complete

I agree with the critics that omission can be thought of as distortion. Still, I do not think that it should undermine the importance of abstraction in science. For the remainder of this section I will put forward some preliminary proposals which show how the omission view can help distinguishing between abstraction and idealization. The first point is that denying the possibility of 'mere omission' altogether is too strong. Phenomena in the world have a very large number of parts and properties and scientists always omit some of them in their experiments and models. Some of these properties do not have an effect on the study. For example, one of the properties of the VI marmots is eye color. The paper does not make any reference to this property, because the scientists did not think that it was relevant for population growth. I think it is safe to say that the property of eye color which was present in the system, was 'merely omitted' from the model.

The upshot is that abstraction and idealization are distinct processes that give rise to different types of phenomena. Therefore the norms that govern these processes should also be different. There is a substantial literature that deals with the methodology and evaluation of idealizations (see for example Giere 1988, Weisberg 2007a). An idealization misrepresents a factor that is considered important for the phenomenon of interest, by adding properties to it or changing some of its properties. For example, scientists may assume that a population is infinite, in order to construct an evolutionary model that is computationally tractable. In order to be successful, the idealized system must be informative about the real system, despite the misrepresentations. This can be achieved if the idealized system is to some extent

isomorphic to its real-world counterpart, or if it is sufficiently similar to it (van Fraassen 1980, Weisberg 2007b).

The case of abstraction is different. Phenomena in nature have many more parts and properties than one can include in an experiment or model. Hence, when scientists abstract they want to preserve only those parts and properties that are relevant for the phenomenon they are studying. These omissions help them make sense of the phenomenon so they can study it. In many cases it might be impossible to study a phenomenon without omitting a large number irrelevant factors. As stated above, when abstracting, scientists aim for 'mere omission'. Therefore, the evaluation of an abstraction should focus on whether the it is a case of 'mere omission' or not. To my knowledge, there is no account that fully specifies a method for the evaluation of abstractions.[5] It is usually left to the discretion of the scientist.

It unlikely that the methods used to evaluate idealizations (such as isomorphism or similarity) can be applied to the evaluation of abstraction. Abstract systems are already very similar to their real-world counterparts, because they are concrete and real. The differences between concrete systems at different levels of abstraction are much more fine-grained than differences between idealized and real systems. Also, an abstract system can be to a large extent isomorphic to a complete system, yet lack a relevant property. For example, an experiment that looked at competition between *T.angustifolia* and *B.fluviatilis*, which focused only on above-ground competition and did not take into account below-ground competition would be isomorphic to the real-world ecosystem,

[5] There are some accounts that outline important aspects of the process of abstraction (for example Jones 2005, Weisberg 2007). Still, these accounts are focused on describing the process of abstraction and do not give a generalized account of how abstractions should be evaluated.

yet it would also be missing relevant aspects of complete system.[6] Thus, it seems that a different method is needed for a full and generalized evaluation of abstraction in science. This account will have to wait for another paper. The purpose of this paper was to show that before any such account is possible, the omission view must be distanced from the material view of abstraction and hence from idealization.

IV. Conclusion: Abstract and Complete

The two notions of abstraction captured by the material view and the omission view respectively, are easily confused. The examples that are usually used to illustrate discussions of abstraction exacerbate the situation, as they are often taken from mathematics and mathematical objects are seen as paradigm examples of abstract objects. While the distinction might not be necessary in mathematics, it is very important for science, especially biology. Failing to distinguish between the two notions undermines the role that abstraction plays in scientific experimentation and modeling, as it is often subsumed under the concept of idealization. Keeping these two concepts separate will give us a more accurate picture of scientific methodology and will help in the formulation of a generalized account for the evaluation of the process of abstraction.

[6] This is because allelopathy affects the uptake of nutrients, which occurs in the roots of plants. However, the effects of competition can be seen by looking at the differences in shoot biomass of the competing plants. Still, without the inclusion of below-ground competition and its effect on root biomass, the cause of competition could be missed. That is, if the scientists had not included the below-ground competition in their experiment, they could have overlooked the importance of allelopathy as the *main* cause of *T.angustifolia's* competitive advantage.

Abstract and Complete

V. References

Brashares, J. S., Werner, J. R. and Sinclair, A. R. E. (2010), Social 'meltdown' in the demise of an island endemic: Allee effects and the Vancouver Island marmot. *Journal of Animal Ecology*, 79: 965–973. doi: 10.1111/j.1365-2656.2010.01711.x

Cartwright, N. "Abstract and Concrete." In *Nature's Capacities and Their Measurement*, 183-230. New York: Oxford University Press, USA, 1994. Callaway, R.

Frigg, R. and Hartmann, S. "Models in Science", *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2009/entries/models-science/>.

Giere, R. N. (1988) Explaining Science: A Cognitive Approach. Chicago: University of Chicago Press.

Godfrey-Smith, P. 2006. The strategy of model-based science. *Biology and Philosophy*, 21: 725–740.

Humphreys, P. "Abstract and Concrete." *Philosophy and Phenomenological Research* 55, no. 1 (1995): 157-61.

Jarchow, M. E. & Cook, B. J. Allelopathy as a mechanism for the invasion of Typha angustifolia. *Plant Ecology* **204**, 113-124 (2009).

Jones, M.R. "Idealization and Abstraction: A Framework." *Idealization XII: Correcting the Model. Idealization and Abstraction in the Sciences* (2005): 173-217.

McMullin, E. "Galilean Idealization." *Studies In History and Philosophy of Science Part A* 16, no. 3 (1985): 247-73.

Odenbaugh, J. "Seeing the Forest and the Trees: Realism About Communities and Ecosystems." *Philosophy of Science* 74, no. 5 (2007): 628-41.

Rosen, Gideon, "Abstract Objects", *The Stanford Encyclopedia of Philosophy (Fall 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/abstract-objects/>.

Weisberg, M. (a) "Three Kinds of Idealization." *Journal of Philosophy* 104, no. 12(2007): 639.

(b) "Who is a Modeler?" *British Journal for the Philosophy of Science* 58, pp. 207-233.

No Levels, No Problems: Downward Causation in Neuroscience

Forthcoming in *Philosophy of Science*

Manuscript, June 2012

Markus I. Eronen
Post-doctoral Researcher
Ruhr-Universität Bochum
Institut für Philosophie II; GA03/150
Universitätsstraße 150
D-44780 Bochum
Germany
Email: Markus.Eronen@rub.de

Abstract

I show that the recent account of levels in neuroscience proposed by Bechtel and Craver is unsatisfactory, since it fails to provide a plausible criterion for being at the same level and is incompatible with Bechtel and Craver's account of downward causation. Furthermore, I argue that no distinct notion of levels is needed for analyzing explanations and causal issues in neuroscience: it is better to rely on more well-defined notions such as composition and scale. One outcome of this is that there is no distinct problem of downward causation.

1. Introduction

The notion of "level" appears in several contexts in philosophy of science. For example, the debates on downward causation, mechanistic explanation, reduction, and emergence are conducted in the framework of levels. However, there is no agreement on the definition of a level, or on the criteria for distinguishing levels.

Craver and Bechtel (2007) have recently presented a theory of "levels of mechanisms", which has gained broad acceptance and is currently the most coherent and promising account of levels. They argue for levels of mechanisms, where the relata are mechanisms at higher levels and their components at lower levels. Importantly, these are not general levels of organization, but identified with regard to a certain mechanism. Craver and Bechtel claim that although levels of mechanisms is certainly not the only sense in which "level" is employed in neuroscience or philosophy, it captures the central sense in which explanations in neuroscience span multiple levels. They also employ this theory of levels to deal with the problem of downward causation, arguing that what appears as downward causation can be explained away as same-level causation that has mechanistically mediated effects.

In this paper, I will (1) show that the mechanistic account of levels is unsatisfying, (2) defend an alternative "deflationary" account of levels, where the notion of level is replaced with the more fundamental notions of composition and scale, and (3) explore the consequences this has for the debate on downward causation. My focus

is on neuroscience and downward causation, but the general arguments I raise against levels apply more broadly.

In the next section, I will briefly present the account of levels of mechanisms. In section 3, I will show that this account fails as a theory of levels, since it does not provide any plausible same-level criterion. In section 4, I argue that we should get rid of the problematic notion of "level" altogether and replace it with notions such as scale and composition, which are far better understood. In section 5, I explore some of the consequences this has for the debate on downward causation.

2. Levels of Mechanisms

In most philosophical theories of levels, the core idea is that levels are *compositional*: wholes are at a higher level than the parts that they are composed of (e.g., Oppenheim and Putnam 1958; Wimsatt 1994; Kim 1999). The mechanistic account of levels retains this basic idea, with one important amendment: the relata are not just wholes and parts; they are *behaving* mechanisms and their *active* components. This means that the higher-level entity is an active mechanism performing some function, and the lower-level entities are components that contribute to the mechanism for this function.

Craver gives the following characterization: "In levels of mechanisms, the relata are behaving mechanisms at higher levels and their components at lower levels. These relata are properly conceived neither as entities nor as activities; rather, they should

be understood as acting entities. The interlevel relationship is as follows: X's Φ-ing is at a lower mechanistic level than Ψ-ing if and only if X's Φ-ing is a component in the mechanism for S's Ψ-ing. Lower-level components are *organized together* to form higher-level components." (Craver 2007, 189)

In a similar vein, albeit in more vague terms, Bechtel writes: "Within a mechanism, the relevant parts are … working parts—the parts that perform the operations that enable the mechanism to realize the phenomenon of interest. … It is the set of working parts that are organized and whose operations are coordinated to realize the phenomenon of interest that constitute a level" (Bechtel 2008, 146).

Craver's (2007, 165-170) main example is the case of spatial memory and LTP (Long Term Potentiation), where he identifies four levels. On the top of the hierarchy, there is the level of *spatial memory*, which involves various types of memory and learning. The level of *spatial map formation* includes the structural and computational properties of various brain regions involved in spatial memory, most importantly the hippocampus. The *cellular-electrophysiological* level includes neurons that depolarize and fire, synapses that undergo LTP, action potentials that propagate, and so on. At the bottom of this hierarchy is the *molecular* level, where we find NMDA and AMPA receptors, Ca^{2+} and Mg^{2+} ions, etc. Entities at each lower level are components in a higher-level mechanism: for example, the hippocampus is an active component in the spatial memory mechanism, synapses are active components in the hippocampal mechanism of memory consolidation, and finally, NMDA receptors are active components of the synaptic mechanism of LTP.

Importantly, Craver and Bechtel emphasize that levels of mechanisms are not general levels of organization in the vein of Oppenheim & Putnam (1958), Churchland & Sejnowski (1992) or Wimsatt (1994). “A consequence of this view is that levels are identified only with respect to a given mechanism; this approach does not support a conception of levels that extend across the natural world” (Bechtel 2007).”How many levels there are, and which levels are included, are questions to be answered on a case-by-case basis by discovering which components at which size scales are explanatorily relevant for a given phenomenon” (Craver 2007, 191).

Bechtel and Craver see this as a point in favor of the mechanistic account of levels, since accounts of general levels of organization are ridden with problems: it makes little sense to compare the “level” of glaciers and pyramidal cells, or black holes and microchips. However, the limitations Bechtel and Craver impose are quite extreme: in the mechanistic framework, it does not make sense to ask whether things that belong to different mechanisms are at the same level or not. We cannot even say that a certain molecule in a hippocampus is at a lower level than the hippocampus, unless the molecule is a component of some hippocampal mechanism (Craver 2007, 191).

Even within one mechanism, things that do not stand in a part-whole relation may not be in a level-relation to each other (see, e.g., Craver 2007, 193). One salient example of this is that there is no sense in which the subcomponents of different components of the mechanism are at the same or different level. For example, a component C1 of mechanism M is at one level lower than M, and a subcomponent S1

of C1 is one level lower than the component C1. Another component C2 of M is also one level lower than the mechanism M, and its subcomponent S2 is one level lower than the component C2. However, according to the mechanistic account, the question whether subcomponents S1 and S2 are at the same or different level makes no sense, since they do not stand in a part-whole relation to each other. I return to this issue in the next section.

To summarize, the key features of this account are the following: (1) Levels are "local" – they are always defined relative to one mechanism and the phenomenon of interest. (2) The relata are mechanisms at higher levels and components or "acting entities" or "working parts" at lower levels. (3) Things are assigned to different levels solely based on the part-whole (or component-mechanism) relation: wholes are at a higher level than their parts; parts are at a lower level than the wholes they belong to. In the next section, I show that these features lead to problems, particularly feature (3).

3. Components, Mechanisms, and Problems

Let us consider the mechanism for phototransduction (the conversion of light signals into electrophysiological information) in the retina. Components in this mechanism include rod and cone cells, which are morphologically and functionally distinct types of cells. However, the phototransduction cascade in both rods and cones involves similar components: G proteins (transducin), cyclic guanosine

monophosphate (cGMP), cGMP-gated ion channels, and so on. The cGMP-gated channels in rods and the same types of channels in cones are subcomponents of *different* components of the mechanisms for light adaptation. They do not stand in a part-whole relation. Hence, according to the mechanistic account, there is no sense in which they are at the same or higher or lower level with regard to each other.

However, this is quite implausible. cGMP-gated ion channels in rods and cGMP-gated ion channels in cones are same types of things with same properties, at the same scale, in the same system, and playing a corresponding role in their respective mechanisms (i.e., they are the same types of "acting entities"). If the mechanistic account implies that there is no sense in which these ion channels are at the same level, something seems to have gone wrong, or at least the levels metaphor is used in a way that is extremely unintuitive (I return to this in Section 4).

Things get even more problematic when we consider subcomponents that are causally interacting with each other. For example, consider synaptic transmission between rod cells and (OFF-type) bipolar cells. In the mechanism for synaptic transmission between these cells, active components of the rod cell include synaptic vesicles, which in turn have glutamate molecules as their subcomponents. The active components of the bipolar cells include (AMPA) glutamate receptors, which have "binding sites" as active components. When the rod cell is firing, the glutamate molecules in the vesicles are released, and they bind to the binding sites of the glutamate receptors.

This means that subcomponents (glutamate molecules) of one component (synaptic vesicles) are causally interacting with subcomponents (binding sites) of a different component (AMPA receptors).[1] Yet, Craver and Bechtel explicitly state that there is no sense in which subcomponents of different components are at the same level. This is not only peculiar, but also in fundamental conflict with Craver and Bechtel's (2007) account of cross-level causation: they explicitly defend the view that there is no cross-level or downward causation – causation is an *intralevel* matter, and effects can be then "mechanistically mediated" upwards or downwards in the mechanism. In other words, being at the same level is a *necessary condition* for causal interaction. However, we have now seen that if we follow Craver and Bechtel's own theory of levels, there are clear cases where there are causal interactions between entities that are *not* at the same level. Thus, there is a fundamental conflict between the mechanistic theory of levels and the mechanistic account of downward causation.[2]

---

[1] This is not an isolated example - Fazekas and Kertesz (2011) have recently pointed out other examples and argued that, quite generally, if the components of a mechanism causally interact, also their subcomponents have to causally interact.

[2] I do not want to discuss the nature of causation here, and my main points hold independently of any particular theory of causation. However, the account of causation most naturally fitting the general framework here would be the interventionist theory of causation (e.g., Woodward 2003), which also Craver (2007) explicitly endorses.

These problems are related to the fact that the mechanistic account gives no satisfactory criterion for determining when things are at the *same* level. According to Craver, there is only a partial answer to this question: "X and S are at the same level of mechanisms only if X and S are components in the same mechanism, X's Φ-ing is not a component in S's Ψ-ing, and S's Ψ-ing is not a component in X's Φ-ing." (2007, 192). In other words, what places two items at the same mechanistic level is that they are in the same mechanism, and neither is a component of the other (Craver 2007, 195).

One way of interpreting this is that if any two components in the mechanism are not in a part-whole relation with each other, they are at the same level. However, this would have some bizarre consequences. Consider components X and S in mechanism M. They are at the same level, since X is not component of S and S is not a component of X. Consider then a subcomponent S1 of S. It is also not a component of X, and X is not a component of S1. Then X and S1 are also at the same level, as well as all the further subcomponents of S1 and all their subcomponents! This would be a rather strange account of the same-level relation.

Supposedly the idea is rather that things that are components in a mechanism but not components in any *intermediate* component are at the same level. For example, rod A is at the same level as rod B, since they are components of the phototransduction mechanism and do not stand in a part-whole relation, but a cGMP-gated ion channel in rod B is not at the same level as rod A, because the cGMP-gated ion channel is a component of rod B, and not a "direct" component of the

phototransduction mechanism. Let us call such components that are components in the mechanism directly and not in virtue of being components in another component *direct components*.

If no further restrictions are added, direct components can include things of radically different sizes with very different causal properties. For example, direct components in the mechanism for light transduction in rod cells include things such as the outer segment of the cell, which has the function of capturing photons and may contain billions of opsin molecules. On the other hand, direct components in the mechanism also include single photons hitting the cell, or Na^+-ions in the cell - these are also not components in any intermediate component of the mechanism. It follows that rod outer segments are at the same level of mechanism as photons or Na^+-ions, even though they differ in scale with a factor of at least 10^7 .

Thus, it seems that the same-level criterion that Craver proposes is both too weak and too strong. It is too weak because it implies that in many cases things that are causally interacting and have very similar properties are *not* at the same level. It is too strong because it implies that in many cases things that are of radically different size and that interact at completely different force or time scales are at the same level. This (1) makes the criterion ineffective for distinguishing between interlevel and intralevel causation, and (2) streches the metaphor of "level" near the breaking point.

4. Levels: A Deflationary Account

The main source for the problems outlined above is that the account of Craver and Bechtel is too limited as a theory of levels. It is not an undue exaggeration to say that the account of levels of mechanisms is in fact an account of mechanistic *composition*: it relies entirely on the component-mechanism relation and simply labels whole mechanisms as being at higher "levels" and their components as being at lower "levels". For this reason, it is difficult to define any reasonable same-level relation in this framework: composition only relates parts and wholes, and not parts with other parts or wholes with other wholes.

My suggestion is, first of all, to take the approach of Craver and Bechtel into its logical conclusion and to deflate the notion of mechanistic levels into simply mechanistic composition. We can simply reinterpret the mechanistic account of levels as an account of mechanistic composition, as long as we strip away the idea of being at the "same" mechanistic level and the related claims about same-level causation. I fully agree with Craver and Bechtel in that explanations in neuroscience refer to robust properties and generalizations throughout the compositional hierarchy – for example, in the explanation for phototransduction we need to consider the 11-cis-retinal molecule changing shape, the rod photoreceptor cell hyperpolarizing, the retinal network computing, the eye converting light to electrophysiological signals, and so on.

However, it is obvious from section 3 that this will not be sufficient as a framework for dealing with issues such as downward causation. Therefore, the second step of

my solution is to take into account the dimension of *scale*, which is largely independent from composition. In his discussion of levels, Craver (2007, ch. 5) acknowledges the importance of size scale, but argues that it is secondary to composition: components cannot be larger that the wholes they are part of, so in this sense the size dimension partly follows the compositional dimension. However, we have also seen above that composition and size often come apart: the direct components of a mechanism can be of radically different sizes, and similarity or difference of size does not imply that entities are in any way compositionally related. Composition and scale are largely independent dimensions (see also Richardson and Stephan 2007; Rueger & McGivern 2010).

The most commonly discussed scale is size scale, but also other scales such as the temporal scale (the speed of interactions) or the force scale (the strength of interactions) may be just as important in understanding complex systems (see, e.g., Simon 1962; Rueger & McGivern 2010). For example, molecular interactions happen at a much faster time scale than interactions between neurons, which are again faster than interactions between brain areas. The force scale is particularly important when considering physical and chemical interactions: for example, the forces binding subatomic particles (quarks) together are much stronger than the forces binding atoms together, which are again stronger than the forces binding molecules together. For the sake of clarity, I focus here mostly on the size scale.

One problem of the mechanistic account of levels was that its same-level relation leads to results that seem arbitrary and unintuitive: for example, there is no sense in

which subcomponents of components are at the same mechanistic level, even when they are same types of things, while entities of radically different sizes can be at the same level. In my view, it is better to get rid of the idea of being at the "same level" altogether, and just to focus on how things are related on different scales (see also Potochnik & McGill 2012). For example, cGMP-gated ion channel are obviously found at the same size (and temporal) scale than cGMP-gated ion channels in cones, while rod outer segments are found at very different size (and temporal) scales than Na^+ ions.

One outcome of analyzing levels in terms of scale and composition is that we no longer need any distinct notion of level. If scale and composition are sufficient for analyzing explanations in neuroscience, the notion of "level" does not add anything to our conceptual toolkit. Explanations in neuroscience are "multilevel" only in the sense that they refer to robust properties and generalizations at various stages in the compositional hierarchy and at different (size) scales.

This approach is also supported by neuroscientific practice. In contrast to what Craver (2007, ch. 5) suggests, levels talk is not very common in neuroscience, neither in journal articles nor in standard textbooks such as Kandel, Jessell and Schwartz (2000) or Purves et al. (2004). In many articles (see, e.g., Malenka & Bear 2004) the term does not come up at all. When it does appear, it is most often referring to levels of processing, such as the different stages of visual information processing (the retina, the LGN, the visual cortex, and so on), which are something

very different from levels of mechanisms, and "levels" only in a metaphorical sense.[3] This supports my point that the notion of level does not pick up any distinct or important category.[4]

If one insists on using the term "level" to refer to stages of composition or to different size scales (or to various other things – scale and composition are merely the senses most relevant in this context), one has to at least make clear in exactly which sense the term is used. However, the danger in this is that other intuitions about levels may creep in – for example, when talking of compositional stages as "levels", one is easily lead to think that things can be at the "same level" of composition.

5. Downward Causation and Levels

I have argued above that the idea of levels is thoroughly problematic, at least in philosophy of neuroscience, and that we should abandon the project of trying to define levels. Let us now turn to the issue of downward or top-down causation that has been traditionally discussed in the framework of levels (e.g., Campbell 1974;

---

[3] Of course, the *word* "level" often comes up in the trivial sense of "luminance level", "level of oxygen", "level of noise", etc.

[4] Ladyman and Ross (2007, 54) reach a similar conclusion in the philosophy of physics.

Emmeche et al. 2000; Kim 1992, 1999; Craver and Bechtel 2007; Kistler 2009).[5] The question is whether higher-level causes can have lower-level effects. In spite of various arguments to the effect that downward causation is not possible, the debate keeps resurfacing, partly because (neuro)scientists often rely on top-down experiments and explanations that seem to imply some kind of downward causation.

As we have seen above, Craver and Bechtel (2007) have proposed a novel solution to the problem of downward causation. They argue that what appears to be downward causation in top-down experiments and elsewhere should be understood as normal same-level causation that has "mechanistically mediated" effects downwards in the mechanism: there is no causation from higher to lower levels or the other way around.

Considering the discussion in the previous two sections, it is clear that the reason why the solution of Craver and Bechtel does not work is that it relies on the distinction between same-level and cross-level causation. We have seen how difficult it is to define the same-level relation, or levels in general, in a coherent and scientifically plausible way. The term "level" does not seem to pick up any distinct

---

[5] In a recent article, Love (2012) discusses top-down causation in terms of levels, but in a way that comes closer to my approach: he argues that there are many different kinds of level-hierarchies and correspondingly many different kinds of top-down causation.

category in neuroscience. For this reason, basing the account of downward causation on the distinction between same-*level* causation (which is supposed to be unproblematic) and cross-*level* causation (which is supposed to be unacceptable) necessarily leads to problems.

One possibility would be to try to reformulate Craver and Bechtel's solution in terms of scale and composition. If we could distinguish between same- and different-"level" causation in terms of scale and composition, perhaps the solution could still work. Unfortunately, this does not seem to be the case. As I have already pointed out in the previous section, composition as such does not involve any same-"level" relation. Regarding (size) scale, the problem is that there is absolutely no reason to restrict causation to things of same or similar size: elephants squash flies, the fission of uranium atoms causes cities to disintegrate, and so on. Therefore, we have to conclude that Craver and Bechtel's approach downward causation is unsatisfactory.

If we abandon the framework of levels and focus on scale and composition, what appears to be downward causation reduces to two categories: (1) Causes that act from the mechanism as a whole to the components of the same mechanism, and (2) causation between entities of different (size) scales. In my view, it is fairly clear that there can be no causation between things that are related by composition (category (1)), since composition is a form of non-causal dependency. It does not seem right to say that, e.g., the retina as a whole causes a rod cell in that retina to fire. On the other hand, as the examples in the previous paragraph show, causation between things of

different size[6] is in principle unproblematic (category (2)). In this way, putative cases of top-down or downward causation can be analyzed away in terms of composition and scale.[7]

One remaining problem for "downward" causation of category (2) is Kim's argument against higher-level causes. It might prima facie seem that getting rid of levels dissolves this problem, since it is often formulated in terms of levels: the argument states that a higher-level property cannot be a genuine cause for a lower-level property, since (due to physical causal closure) the lower-level property already has a sufficient lower-level cause (see, e.g., Kim 1992; 1999). However, the idea of "levels" is not essential in Kim's argument: what is at issue there is the tension created by two competing (and non-causally correlated) causes for the same effect. Without the framework of levels, the argument does not disappear, but turns into the general causal exclusion argument (see, e.g., Kim (2002) Bennett (2008) for more).

---

[6] Whether the same holds for other scales, such as the temporal or the force scale, is an open question that goes beyond the scope of this paper.

[7] One way of interpreting Craver and Bechtel (2007) is that their main point is quite similar, namely that apparent causation from parts to wholes or wholes to parts can be analyzed away in terms of normal causal relations. If this is the case, it is unfortunate that the theory of levels and the distinction between "same-level" and "different-level" causation is so prominent in the paper, since this makes the account unnecessarily complex and confusing.

What Craver and Bechtel (2007) are considering, and what I have discussed in this section, is the intelligibility of causes acting from higher to lower levels. I have argued that downward causation is not intelligible in the sense of causation from a mechanism as a whole to the parts of that same mechanism, but causation from higher to lower scales is as such unproblematic. There may be real problems related to causation in neuroscience, such as the causal exclusion problem, but there is no distinct problem of downward causation.

6. Conclusions

In this paper, I have argued that the account of "levels of mechanisms" is unsatisfactory as a theory of levels, since it does not include a plausible same-level relation, leads to extremely unintuitive results, and is in conflict with the account of downward causation proposed by Craver and Bechtel. Generally speaking, there seems to be no need for a distinct notion or theory of levels in philosophy of mind or neuroscience; it is better to rely on more familiar and well-defined notions such as scale and composition. With this approach, apparent cases of downward causation can be analyzed away.

References

Bechtel, William. 2008. *Mental Mechanisms. Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bennett, Karen. 2008. "Exclusion again." In *Being Reduced*, ed. J. Hohwy and J. Kallestrup, 280-305. Oxford: Oxford University Press.

Campbell, Donald T. 1974. "'Downward Causation' in Hierarchically Organised Biological Systems". In *Studies in the Philosophy of Biology*, ed. F. Ayala and T. Dobzhansky, 179-186. Berkeley, Los Angeles: University of California Press.

Churchland, Patricia S., and Terence J. Sejnowski. 1992. *The Computational Brain*. Cambridge, MA: MIT Press.

Craver, Carl F. 2007. *Explaining the Brain*. Oxford: Oxford University Press.

Craver, Carl F., and William Bechtel. 2007. "Top-down causation without top-down causes." *Biology & Philosophy* 20: 715-734.

Emmeche, Claus, Simo Køppe, and Frederik Stjernfelt. 2000. "Levels, Emergence, and Three Versions of Downward Causation." In *Downward Causation. Minds, Bodies and Matter*, ed. Peter B. Andersen, Claus Emmeche, Niels O. Finnemann, and Peder V. Christiansen. Århus: Aarhus University Press.

Fazekas, Peter, and Gergely Kertész. 2011. "Causation at different levels: tracking the commitments of mechanistic explanations." *Biology & Philosophy* 26: 365-383.

Kandel, Eric R., James H. Schwartz, and Thomas Jessell (2000). *Principles of Neural Science (4th Edition).* New York: McGraw-Hill.

Kim, Jaegwon. 1992. "'Downward Causation' in Emergentism and Nonreductive Physicalism." In *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*, ed. Ansgar Beckermann, Hans Flohr, and Jaegwon Kim, 119-138. Berlin: Walter de Gruyter.

Kim, Jaegwon. 1999. "Making Sense of Emergence." *Philosophical Studies* 95: 3-36.

Kim, Jaegwon. 2002. "Mental Causation and Consciousness: The Two Mind-Body Problems for the Physicalist." In *Physicalism and Its Discontents*, ed. Carl Gillett and Barry Loewer, 271-283. Cambridge: Cambridge University Press.

Kistler, Max. 2009. "Mechanisms and Downward Causation." *Philosophical Psychology* 22: 595-609.

Ladyman, James, and Don Ross. 2007. *Every Thing Must Go: Metaphysics Naturalised*. Oxford: Oxford University Press.

Love, Alan C. 2012. "Hierarchy, causation and explanation: ubiquity, locality and pluralism." *Interface Focus* 2: 115-125.

Malenka, Robert C., and Mark F. Bear. 2004. "LTP and LTD: An Embarrassment of Riches." *Neuron* 44: 5-21.

Oppenheim, Paul, and Hilary Putnam. 1958. "Unity of science as a working hypothesis." *Minnesota Studies in the Philosophy of Science* 2: 3-36.

Potochnik, Angela, and Brian McGill. 2012. "The Limitations of Hierarchical Organization." *Philosophy of Science* 79: 120-140.

Purves, Dale, George Augustine, David Fitzpatrick, William Hall, Anthony-Samuel LaMantia, James McNamara, and Leonard E. White, eds. 2008. *Neuroscience*. Sunderland, MA: Sinauer.

Richardson, Robert C., and Achim Stephan. 2007. "Mechanism and mechanical explanation in systems biology." In *Systems Biology: Philosophical Foundations*, ed. F. C. Boogerd, F. J. Bruggeman, J. S. Hofmeyr, and H. V. Westerhoff, 123-144. Amsterdam: Elsevier.

Rueger, Alexander & Patrick McGivern. 2010. "Hierarchies and levels of reality." *Synthese* 176: 379-397.

Simon, Herbert A. 1962. "The Architecture of Complexity." *Proceedings of the American Philosophical Society* 106: 467-482.

Wimsatt, William C. 1994. "The ontology of complex systems: levels of organization, perspectives, and causal thickets." *Canadian Journal of Philosophy* S20: 207-274.

Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Philosophy of Science

The stem cell uncertainty principle

--Manuscript Draft--

Manuscript Number:	
Full Title:	The stem cell uncertainty principle
Article Type:	PSA 2012 Contributed Paper
Keywords:	Experiment, Models, Stem cells, Technology, Uncertainty
Corresponding Author:	Melinda Bonnie Fagan, Ph.D. Rice University Houston, TX UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Rice University
Corresponding Author's Secondary Institution:	
First Author:	Melinda Bonnie Fagan, Ph.D.
First Author Secondary Information:	
Order of Authors:	Melinda Bonnie Fagan, Ph.D.
Order of Authors Secondary Information:	
Abstract:	Stem cells are defined by capacities for both self-renewal and differentiation. Many different entities satisfy this working definition. Explicating the stem cell concept reveals it to be relational and relative, requiring contextualization by a cell lineage, organismal source, cell environments, and traits of interest. Not only are stem cell capacities relative to an experimental context; the stem cell concept imposes evidential constraints on the interpretation of experimental results. In consequence, claims about stem cell capacities are inherently uncertain. I discuss the implications of this result for progress in stem cell research and its public understanding.

Manuscript

1. Introduction

Stem cells are defined as cells that can give rise to more cells like themselves, as well as more specialized, or differentiated, cells.[1] These two cellular processes are termed, respectively, self-renewal and differentiation. A striking feature of stem cell biology is the sheer variety of stem cells: adult, embryonic, pluripotent, induced, neural, muscle, skin, blood, etc. This diversity is exploited in political debates over stem cell funding, and complicates public discussions about stem cells and their therapeutic promise. Stem cells derived from human embryos are cast as ethically dubious alternatives to so-called "adult stem cells" or, more recently "induced pluripotent stem cells."[2] A variety of "stem cell therapies" are touted by medical professionals – some backed by solid evidence, some experimental, and some purely "snake oil."[3] The multiplicity of stem cells, complexity of techniques and terminology, and the passionate nature of debate surrounding their source and potential is such that in some quarters, "the traditional notion of stem cells as a clearly defined class of intrinsically stable biological objects that can be isolated and purified, has begun to give way… the 'stem cell' becomes a fleeting, ephemeral and mythical entity" (Brown et al 2006, 339-343).

To distinguish reasonable hopes from misleading hype, it is necessary to clarify the stem cell concept and its application in various contexts. Philosophers of science have a distinctive role to play here. Bioethicists have approached stem cells as a human reproductive technology, framing debates in terms of moral status, personhood, life and human identity. But this approach

[1] See Melton and Cowan (2009, xxiv), Ramelho-Santos and Willenbring (2007, 35), the 2011 US National Institutes of Health stem cell information page, and the 2011 "Glossary" of the European Stem Cell network. For history of the term, see Maienschein (2003), Shostak (2006).
[2] This 'oppositional' stance made possible the August 2010 injunction on federally-funded embryonic stem cell research in the US, which was imposed because competition for funds allegedly harmed adult stem cell researchers.
[3] See 'About stem cell treatment' at http://www.isscr.org/.

does not fully engage stem cell science, focusing instead on the fragment that manipulates human embryos. This paper argues that the roots of stem cell controversy are not solely in ethics, but also the core concepts and methods of stem cell researchers. I show that pluralism about stem cells, and disagreement about their potential, has conceptual and evidential grounds. This situation gives rise to a deep evidential challenge: the "stem cell uncertainty principle."[4] When clearly stated, this principle makes explicit the uncertainty inherent in the basic stem cell concept. Its constraints have important implications for progress in stem cell research, as well as public understanding of this science.

Section 2 explicates the general stem cell concept, focusing on processes of self-renewal and differentiation. This analysis reveals the key variables and parameters that must be specified for the concept to apply in actual cases; that is, to classify cells (singly or in populations) as stem cells. Section 3 summarizes the core experimental method for identifying stem cells, and shows how it dovetails with the general concept. Stem cell experiments specify the key variables and parameters for particular cases. The evidential challenge posed by these experiments is examined in Section 4. Briefly: stem cell capacities are realized only in descendants. So an individual stem cell can be identified only retrospectively; stem cell researchers literally don't know what they've got until it's gone. The problem cannot be avoided by focusing on cell populations or inventing new techniques. Section 5 considers the implications of this result, and offers suggestions for how stem cell research can progress given its evidential constraints. Section 6 summarizes the conclusions and indicates their broader significance.

Some basic tenets of cell theory are assumed throughout. Every organism begins as a single cell, which, in multicellular organisms, gives rise to all the body's cells. Cells reproduce

[4] This term is from Nadir (2006).

by binary division.[5] The life of a cell begins with a division event and ends with either a second division event yielding two offspring, or cell death yielding no offspring. Generations of cells linked by reproductive division form a lineage. Self-renewal is cell reproduction in which parent and offspring resemble one another. Differentiation, along with growth, is the core phenomenon of development: the process by which parts of a developing organism acquire diverse, specialized traits over time. These premises provide the background for further clarification of the stem cell concept.

2. Stem cell concept

Stem cells are defined as cells capable of both self-renewal and differentiation. The simplest way to conceptualize a stem cell is in terms of a cell division event that includes both processes: one cell like the parent, the other more specialized (Figure 1a). But this simple model does not capture the stem cell concept. No two cells are the same or different in every respect. At minimum, the cells involved in a division event (one parent and two offspring) differ in position and intercellular relations, and share some material parts, including DNA sequences. Comparisons that determine 'stemness' must be made relative to some set of characters, such as size, shape, concentration of a particular molecule, etc. Given a set of characters C={x, y, z…n}, values within and across cell generations can be compared, to determine relations of sameness and difference among cells in a lineage (Figure 1b).

[FIGURE 1]

[5] There are two modes of cell division: mitosis and meiosis. In mitosis, the genome replicates once before the cell divides. In meiosis, the genome replicates once, but two rounds of cell division follow, yielding four offspring cells with half the complement of DNA. Stem cell phenomena involve mitosis, so the term "cell division" here refers to that mode only.

2.1 Stem cell capacities

The above is still insufficient to define self-renewal and differentiation, which have temporal as well as comparative aspects. The dynamic aspect of self-renewal is conceived as the number of division cycles in which parent and offspring cells are the same with respect to some set of characters C (Figure 2a).[6] Differentiation involves change within a cell lineage over a time interval t_2-t_1. The simplest way to conceive of cellular change is in terms of a single cell with some character X (e.g., shape or size), which has value x_1 at time t_1 and x_2 at a later time t_2. But not every such change counts as differentiation. A cell that changes character value from x_1 to x_2 thereby differentiates only if the change is 'directed' in at least one of two ways: toward more specialization or greater diversity. These two 'directions' correspond to two kinds of comparison: between cells of a developing lineage, and between developing and mature cells (Figure 2b). The former become more heterogeneous over time, differentiating from one another. More precisely, cells in lineage L diversify over time interval t_2-t_1, relative to a set of characters C, if and only if values of C vary more at t_2 than t_1. The second comparison is between cells that have completed development and those that have not. The diverse cells composing the body of a fully-developed organism are classified according to typologies that may extend to hundreds of cell types. Each of the latter is defined by a cluster of character values, C_m. A cell specializes over time interval t_1-t_2 just in case its character values are more similar to C_m at t_2 than at t_1.[7] The relevant set of characters is determined primarily by attributes of mature cells that are the end-points of the process.

[6] Cell cycle rate converts this to calendar time; in practice both measures are used.
[7] In many cases, however, there is not one cell fate to consider, but a whole array, each with a characteristic complex of traits (C_{m1}, C_{m2}…C_{mn}). So, in general, a cell specializes over t_1-t_2 if its

[FIGURE 2]

These considerations support the following characterizations of the reproductive processes that define stem cells:

(SR) Self-renewal occurs within cell lineage L relative to a set of characters C for duration τ, if and only if offspring cells have the same values for those characters as the parent cell(s).

(DF) Differentiation occurs within cell lineage L during interval t_1-t_2 if and only if character values of some cells in L change such that (i) cells of L at t_2 vary more with respect to characters C than at t_1, or (ii) cells of L at t_2 have traits more similar to traits C_m of mature cell type(s) than at t_1.

Putting the two together yields a general definition of 'stem cell': a stem cell is the unique stem of a branching structure organized by SR and DF, such that each branch terminates in exactly one mature cell type (Figure 2c). This minimal, abstract model[8] structurally defines a stem cell by position in a cell hierarchy organized by reproductive relations.

2.2 Parameters

---

traits are more similar to some C_m at t_2 than at t_1. The attributes of specialized mature cells are so various that it is awkward to conceive them as values of a single set of characters. A cell can become more similar to an adult cell type either by changing values of a set of characters C (x_1 to x_2), or by changing its set of characters (C to C').

[8] 'Model' here is used in Giere's sense (1988).

This minimal model covers every case of stem cells. But on its own, it entails no predictions about cell phenomena. Representational assumptions are needed to connect its objects to biological targets. Three different representational assumptions are prevalent in stem cell biology today, interpreting the model's objects as: (i) single cells undergoing division; (ii) reproductively-related cell populations with statistical properties; or (iii) reproductively-related cell types. In addition, applying the minimal stem cell model requires specification of its key parameters and variables: temporal duration and characters of interest. Whether a given cell counts as a stem cell depends, in part, on how these parameters are specified. Table 1 summarizes the parameters associated with the major stem cell types in use today.

[TABLE 1]

In general, the shorter the duration of interest, the lower the bar to qualify as a stem cell. Most stem cell research is concerned with longer intervals, so the bar to qualify as a stem cell is higher. But there is no absolute threshold. What counts as a stem cell varies with the temporal duration of interest. Another variable is number of terminating branches in the cell lineage hierarchy. Termini of these branches are cell fates, each distinguished by a "signature" cluster of character values, C_m. The more terminating branches emanate from a cell, the greater its developmental potential. The maximum possible developmental potential is totipotency: the capacity to produce an entire organism (and, in mammals, extra-embryonic tissues) via cell division and differentiation. In animals, this capacity is limited to the fertilized egg and products of early cell divisions. In the late-19^{th}/early-20^{th} century, such cells were referred to as stem cells, but terminology has since shifted. The maximum developmental potential for stem cells in

the contemporary sense is pluripotency: ability to produce all (major) cell types of an adult organism. Somewhat more restricted stem cells are multipotent: able to produce some, but not all, mature cell types. Stem cells that can give rise to only a few mature cell types are oligopotent. The minimum differentiation potential is unipotency: the capacity to produce a single cell type. This classification of potencies, though imprecise, provides a convenient framework for comparing stem cells associated with different cell traits and fates (Table 1).

Finally, applying the abstract model requires criteria to judge cells the same or different with respect to a set of characters. Our only access to cells is via technologies that visualize, track, and measure them. So character values attributed to cells are very closely associated with methods of detection. Cells in adult organisms are distinguished by morphological, histological, and functional criteria, which figure prominently in typologies. Undifferentiated cells are often characterized negatively, as lacking these traits. Cell traits, fates, and technologies for distinguishing them are all closely entwined. Specifying criteria for cell character values to count as the same or different amounts to specifying a set of methods for measuring those characters. This brings us to concrete experiments that identify stem cells.

3. Methods

Methods for identifying stem cells share a basic structure of three stages (Figure 3a). The starting point is a multicellular organism, the source of cells. From this source, cells are extracted and values of some of their characters measured. These cells (or a sample thereof) are then manipulated so as realize capacities for self-renewal and differentiation. Each experiment involves two manipulations. In the first, cells are removed from their original context (a multicellular organism) and placed in a new environment in which their traits can be measured.

Second, measured cells are transferred to yet another environmental context, which allows stem cell capacities to be realized. Finally, the amount of self-renewal and differentiation is measured. Stem cell experiments[9] thus consist of two manipulations, each followed by measurement, of cells from an organismal source.

[FIGURE 3]

This basic method identifies stem cells by three sets of characters: of organismal source, of extracted cells, and of progeny cells (Figure 3b). The characters included in the first and third sets are standardized and robust across a wide range of experiments. For organismal source, these characters are species, developmental stage, and tissue or position within the organism.[10] Values of these characters are determined by choice of materials for an experiment: mouse or human; embryonic or adult; blood, muscle or a quadrant of the early embryo. Values for the other two sets of characters are measured during an experiment. For progeny cells, characters included are those of mature cell types: morphology, expression of specific genes and proteins, and function within an organism. Exactly which characters comprise the set depends on the type of differentiated cells expected. For blood cells, the relevant characters are associated with immune function; for neurons, electrochemical function; for germ cells, morphological and genetic traits of gametes. Though the set of characters varies across experiments, for any particular experiment the characters of interest are established in advance: part of the standard set

[9] Stem cell biology includes many kinds of experiment. For brevity, I refer to experiments that aim to isolate and characterize stem cells as 'stem cell experiments.' But this should not be interpreted as exhaustive of experiments in the field.
[10] Another frequently-used organismal character is genotype or strain.

of morphological, biochemical and functional traits used to classify cells in multicellular organisms.

In contrast, there are no such pre-established criteria for inclusion in the set of characters of extracted cells – i.e., presumptive stem cells. These characters vary widely across experiments, shifting rapidly in response to technical innovations and new results within the field. Yet measurement of their values is the linchpin of stem cell experiments. Experiments aimed at isolating and characterizing stem cells succeed just in case they reveal the "signature" traits of stem cells from a given source. Relations among values of these variables map features of organismal source and differentiated descendants onto a 'stem cell signature,' entailing many predictions. A predictive model of this sort would describe robust relations between the values of variable characters in these three domains. We do not yet have such a model, however; 'mapping' relations among source, signature, and progeny are largely unknown, even for the best-understood stem cells. Indeed, the 'stem cell signatures' we have are at best provisional. An important goal of stem cell research is to flesh out this speculative sketch. But here the stem cell concept itself poses a serious challenge.

4. Uncertainty

Stem cell experiments involve two sets of measurements, both of which provide data about characters of single cells. But no single cell persists through both sets of measurements. Cells reproduce by division, so descendents and ancestors cannot co-exist. The second set of measurements is of cells descended from those measured in the first. Self-renewal and differentiation potential are measured after realization of these capacities in controlled environments: the second set of measurements. A single stem cell, therefore, can be identified

only retrospectively. At the single-cell level, stem cell researchers literally don't know what they've got until it's gone.

There are three distinct evidential problems here. First, self-renewal and differentiation potential cannot both be measured for a single cell. To determine a cell's differentiation potential, that cell is placed in an environment conducive to differentiation, and its descendants measured. To determine a cell's self-renewal ability, the cell is placed in an environment that is conducive to cell division without differentiation, and its descendants measured. It is not possible to perform both experiments on a single cell. Since stem cells are defined as having both capacities, stem cells cannot be identified at the single-cell level. Second, the capacity for self-renewal cannot be decisively established for any stem cell. An offspring cell with the same capacities as a stem cell parent has the same potential for differentiation and for self-renewal. Even if both could be measured for a single cell (which they cannot), it is the offspring of the offspring cell that indicates the latter's capacities. The relevant data are always one generation in the future. Experimental proof that a single cell is capable of self-renewal is infinitely-deferred. Third, in any experiment, differentiation potential is realized in a range of (highly artificial) environments. But these data cannot tell us what a cell's descendants would be like in a different range of environments – in particular, physiological contexts. There is, inevitably, an evidential gap between a cell's capacities, unmanipulated by experiment, and their realization in specific, highly artificial, contexts. For all three reasons, claims that any single cell is a stem cell are inevitably uncertain. This uncertainty admits diverse, even arbitrary, operational criteria for self-renewal, and underpins perennial debate over the extent of differentiation potential in stem cells from adult organisms.

These evidential limitations of stem cell experiments have been likened to the Heisenberg uncertainty principle, which states that a particle's mass and velocity cannot be simultaneously measured. In physics, the procedure used to determine the value of one alters the value of the other. The analogy suggests that measurement itself is the problem; e.g., "…we cannot determine both the function of a cell and its functional potential…[because] our determination of a cell's function at a given point in time interferes with an accurate determination of its developmental potential" (Nadir 2006, 489), and we cannot rule out the possibility that "the investigator might be forcing the stem-cell phenotype on the population being studied" (Zipori 2004, 876). But for stem cell biology, the problem is not measurement of cells per se, but their transfer to different environmental contexts. Stem cell capacities are realized and measured in cells descended from 'candidate' stem cells, in different environments (for differentiation potential). Potten and Loeffler (1990) articulate the issues incisively:

> The main attributes of stem cells relate to their potential in the future. These can only effectively be studied by placing the cell, or cells, in a situation where they have the opportunity to express their potential. Here we find ourselves in a circular situation; in order to answer the question whether a cell is a stem cell we have to alter its circumstances and in so doing inevitably lose the original cell and in addition we may see only a limited spectrum of responses… Therefore it might be an impossible task to determine the status of a single stem cell without changing it. Instead one would have to be satisfied with making probability statements based on measurements of populations (1009).

It might seem that stem cell biologists can avoid these problems by shifting their focus to cell populations. Representational assumptions (ii-iii) allow for exactly this (see §2 above). Two kinds of model, stochastic and compartmental, yield hypotheses about stem cell

populations.[11] But experimental support for these hypotheses depends on hypotheses about single stem cell traits. Here I address stochastic population models only; an analogous argument can be made for compartment models.[12] Stochastic population models of stem cells are based on the following assumptions. Any population of cells experiences some number n divisions over a period of time τ, such that the population grows, diminishes, or remains constant in size. Any dividing cell in the population has a certain probability of undergoing each of three kinds of division: both offspring like the parent (p), one offspring like the parent (r) or no offspring like the parent (q), where $p + r + q = 1$. Relations among p, r, and q values entail general predictions about cell population size (growth, decrease, or "steady-state"), and equations that predict mean and standard deviation in population size, probability of stem cell extinction, and features of steady-state populations are derived.[13] In these equations, p is the fundamental parameter. Testable predictions require that its value be estimated. This is done by estimating the coefficient of variation for stem cell number in populations of the same age produced by division from a single founding stem cell. The data required for such an estimate are numbers of stem cells in replicate colonies, each originating from a single stem cell.

Given such an estimate, a stochastic stem cell model predicts features of cell population kinetics, which can then be compared with experimental data. But the hypothesis thereby tested is not that 'founder' cells are stem cells. Rather, it is that stem cell population size is regulated so as to yield predictable population-level results from randomly-distributed single-cell capacities. Testing this hypothesis requires identifying stem cell populations. Stochastic models make predictions, given the assumption that 'founding elements' are stem cells. All these

[11] Terms from Loeffler and Potten (1997).
[12] [reference removed for blind review]
[13] Details in Vogel et al (1969).

predictions hinge on estimation of the fundamental parameter p, the probability that a stem cell undergoes self-renewal. This parameter is estimated from the pattern of variation in a set of replicate colonies, initiated by a single "stem element." But in order for experiments to be replicates, all the stem elements for the set of colonies must be assigned the same probability values for p and (1-p); i.e., the same capacities for self-renewal and differentiation. So experimental test of a stochastic stem cell model depends on the assumption that the cell population measured is homogeneous with respect to these characters. This is exactly the evidence that the stem cell uncertainty principle ensures we cannot get. Stochastic population-level stem cell models therefore do not avoid the evidential challenge above.

To sum up: stem cell experiments, no matter how technically advanced at tracking and measuring single cells, cannot resolve stem cell capacities at the single-cell level. This is because we cannot directly measure a single cell's capacity for self-renewal or differentiation, separately or together. To measure both self-renewal and differentiation potential for a single cell, and to elicit the full range of a cell's potential, multiple 'copies' of that cell are needed - a homogeneous cell population of candidate stem cells. Thoroughgoing focus on cell populations cannot get around this problem, since evidence for population-level models of stem cells also depends on the assumption of a homogeneous 'founder' stem cell population. The 'uncertainty principle' is an unavoidable evidential constraint for stem cell biology.

5. Progress

How, then, should stem cell biologists proceed? In practice, the dominant strategy is to adopt a 'single-cell standard;' that is, to assess progress not in terms of hypotheses, but experimental methods. Better experimental methods improve our access to single cells. Current "gold

standards" for stem cell experiments are articulated in exactly these terms. These standards are implemented somewhat differently for stem cells with different potencies. For 'tissue-specific' stem cells, the gold standard is a single-cell transplant leading to long-term reconstitution of an animal's tissue or organ. An ideal pluripotent stem cell line behaves as a single cell, exhibiting the same traits in the same culture environment, so self-renewal or differentiation capacities can be realized on demand.[14] But across the entire field, technologies that enhance our ability to isolate or track single cells are quickly adopted and reported as advances.[15] Post-genomic and micro-imaging technologies are increasingly important in stem cell biology, for this reason. But the single-cell standard dates back to post-WWII experiments with cultured cells and transplantable tumors in inbred mice. The first method for measuring stem cells was announced as "a direct method of assay for [mouse bone marrow] cells with a single-cell technique" (Till and McCulloch 1961, 213).

This approach is evidentially well-founded. The single-cell standard, applied across many stem cell types (i.e., experimental contexts), supports the assumption of homogeneity on which all stem cell models depend. An experiment that meets the standard begins with a single cell in a controlled environment, with all relevant signals that could impact the cell taken into account. If all other cell reproduction in this environment is blocked, or products of the founding cell can be distinguished from all other cells, then results reflect the reproductive output of a single starting cell, and no others. Measured stem cell capacities are then unambiguously attributed to that cell in that environment. Technologies that track a single cell's reproductive output over time, combined with techniques that measure character values of single cells, can

[14] "Gold standards" from Fundamentals of Stem Cell Biology (Cowan and Melton 2009) and the International Stem Cell Initiative's characterization of hESC lines (Adewumi et al 2007).
[15] For recent examples, see special issues of Nature Reviews Genetics (April 2011) and Nature Cell Biology (May 2011).

yield data of this sort. In this way, technical innovations guided by the single-cell standard can bolster evidence for stem cell models – but only relative to the environment in which stem cell capacities are realized. More general results are obtained from replicate experiments using a range of environments. If the same environment tends to elicit self-renewal of the same duration and/or differentiation into the same cell types, while different environments reliably yield different results, this indicates that the cell population from which replicates are drawn is homogeneous with respect to stem cell capacities. Of course, populations homogeneous with respect to one set of character values need not be homogeneous with respect to others. But sorting cells into populations homogeneous for many measurable traits is the best we can do, since stem cells cannot be identified in advance.

So the 'stem cell uncertainty principle' does not block progress in stem cell research. But, since the possibility of heterogeneity in stem cell capacities cannot be completely ruled out, hypotheses about stem cells can never be fully and decisively established. Stem cell experiments can provide good evidence for hypotheses at the single-cell level, but only relative to the set of characters used to specify a homogeneous sub-population. As new cell traits are discovered and made accessible to measurement, the assumption of homogeneity must be continually reassessed and revised. All substantive models of stem cells are therefore necessarily provisional, and become obsolete when new characters and environments are introduced. This evidential constraint necessitates a mode of collaboration in stem cell research that gives the lie to the idea that the field is essentially a competition of models and methods in a 'race to the cure.' Improved single-cell methods applied to all available stem cell types gives rise to a whole constellation, or network, of improved models. In this way, guided by experiment, the entire field moves forward together.

6. Conclusions

The basic stem cell concept is relational and relative. So stem cells are not defined absolutely, but relative to an organismal source, cell lineage, environments, traits and a temporal duration of interest. Experimental methods for identifying stem cells specify these parameters. In any actual case, therefore, stem cells must be understood in terms of experimental methods used to identify them. The stem cell uncertainty principle imposes evidential constraints on these methods, however. Several consequences follow. First, all stem cell claims are provisional, dependent on an assumption of cell homogeneity that must be continually reassessed as research moves forward. Second, stem cell pluralism is not a symptom of incomplete understanding, but follows from the general stem cell concept. Claims about stem cells based on different elaborations of the basic model do not conflict. The diversity of stem cells should not be a source of contention, but a positive resource for inquiry. Finally, technical innovations that increase experimenters' ability to measure and track single cells can bring about a situation in which experiments can provide strong evidence for hypotheses about stem cells. 'Single-cell' technologies are thus an important form of progress in stem cell biology, with evidential significance.

Acknowledgements

[removed for blind review]

References

Adewumi, O., et al. (2007), "Characterization of human embryonic stem cell lines by the International Stem Cell Initiative", Nature Biotechnology 25: 803-816.

Brown, N., Kraft, A., and Martin, P. (2006), "The promissory pasts of blood stem cells", BioSocieties 1: 329-348.

Giere, R. (1988), Explaining Science. Chicago: Chicago University Press.

Loeffler, M., and Potten, C. S. (1997), "Stem cells and cellular pedigrees", in: Potten, C. S. (ed.) Stem Cells. London: Academic Press, 1-27.

Maienschein, J. (2003), Whose view of life? Cambridge, MA: Harvard University Press.

Melton, D.A., and Cowan, C. (2009) "Stemness: definitions, criteria, and standards", in: Lanza, et al (eds.) Essentials of Stem Biology, 2nd edition. San Diego, CA: Academic Press, pp. xxii-xxix.

Nadir, A. (2006), "From the atom to the cell", Stem Cells and Development 15: 488-491.

Potten, C. S., and Loeffler, M. (1990), "Stem cells: attributes, cycles, spirals, pitfalls and uncertainties", Development 110: 1001-1020.

Ramalho-Santos, M., and Willenbring, H. (2007), "On the origin of the term 'stem cell'", Cell Stem Cell 1: 35-38.

Shostak, S. (2006), "(Re)defining stem cells", BioEssays 28: 301-308.

Till, J.E. and McCulloch, E.A. (1961), "A direct measurement of the radiation sensitivity of normal mouse bone marrow cells", Radiation Research 14: 213-222.

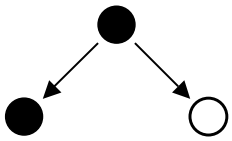
Vogel, H. et al. (1969), "Stochastic development of stem cells", Journal of Theoretical Biology 22: 249-270.

Zipori, D. (2004), "The nature of stem cells", Nature Reviews Genetics 5: 873-878.

Figure
Click here to download Figure: Figures_PSA2012.pdf

Figure 1 Simple stem cell model: (a) single cell, (b) cell population.

A.



B.

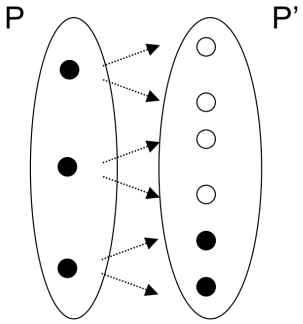


Figure 2 The stem cell concept: (a) self-renewal, (b) differentiation, (c) both. Arrows represent cell reproductive processes, variables represent key parameters (see text).

A.

traits $x_1, y_1, \ldots z_1$
C: {x,y, z…n}
$1 \leq t \leq \infty$ cell cycles

B.

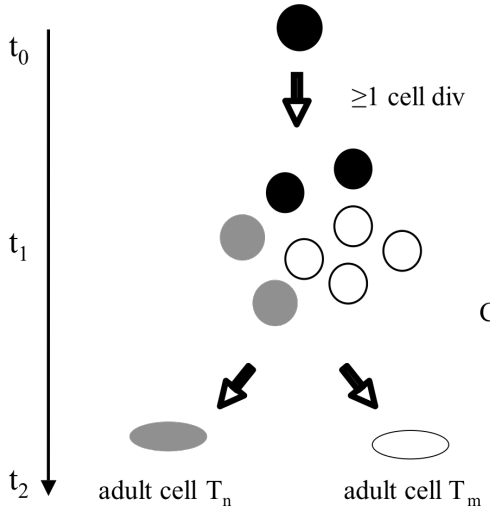


Figure 2, cont.

C.

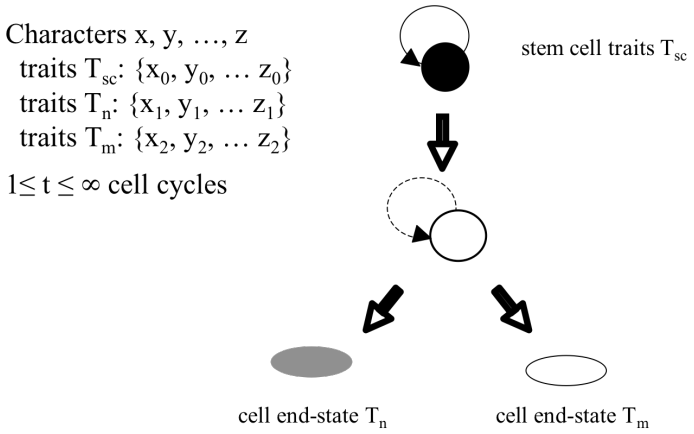
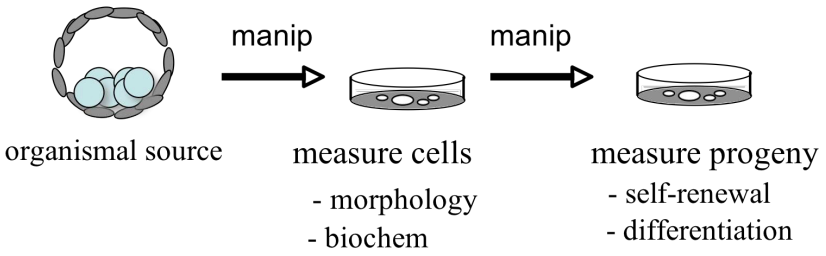


Table 1 Stem cells, classified in terms of the general model and its key parameters. (For simplicity, time intervals are left approximate and only characters are indicated, not specific values. The latter are diverse; 'various' indicates that no standard is widely-accepted for a stem cell type.)

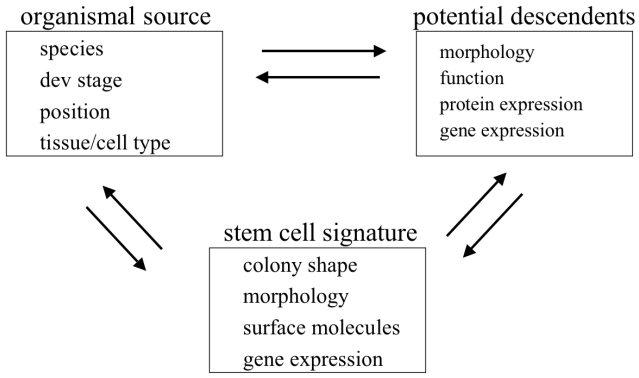
Stem cell	Characters	Time interval/ duration	Potency	Source
ESC	shape, size, cell surface markers, gene expression	indefinite (>50 cycles)	pluripotent	early embryo inner cell mass
HSC	various	various (wks-decades)	multipotent	bone marrow, cord blood, peripheral blood
NSC	morphology, cell surface markers, nerve function	months-years	oligopotent	brain (adult and embryonic)
iPSC	shape, size, cell surface markers, gene expression	months-years	pluripotent	differentiated cells (various tissues)
epiSC	shape, size, cell surface markers, gene expression	months-years	pluripotent	early embryo inner cell mass
GSC	shape, size, cell surface markers, gene expression	months-years	pluripotent	genital ridge (embryo)
CSC	various	?	?	cancer (leukemia)
EC	shape, size, cell surface markers	weeks-months	pluripotent	cancer (teratocarcinoma)
epiderm	morphology, cell surface markers	years	unipotent	skin
hair	morphology, cell surface markers	years	unipotent	follicle

Figure 3 Basic design of stem cell experiments: (a) experimental procedure, (b) results.

A.



B.



TRUE Is False and Why It Matters

Robert William Fischer

Texas State University-San Marcos

1. Introduction

The proponents of inference to the best explanation (IBE) are willing to move from a judgment about the quality of an explanation to a judgment about its probability. In other words, they sanction inferences that have this form:

P1: Facts $f_1 - f_n$ obtain.
P2: If true, hypothesis *h* would offer a better explanation of $f_1 - f_n$ than would any competing hypothesis.
C: So, probably, *h* is true.[1]

This formulation raises an obvious question: for a given hypothesis, in what sense is it alleged to offer a 'better' explanation of $f_1 - f_n$? The standard answer is that the hypothesis has a higher score on the explanatory virtues: conservatism, modesty, simplicity, generality, and predictive power.[2] But this answer appears to be problematic. Here Bas van Fraassen's objection to it:

> Judgments of simplicity and explanatory power are the intuitive and natural vehicle for expressing our epistemic appraisal. [But these] are specifically human concerns, a function of our interests and pleasures, which make some theories more valuable or appealing to us than others. Values of this sort […] provide reasons for using a theory, or contemplating it, whether or not we think it true, and cannot rationally guide our epistemic attitudes and decisions. For example, if it matters more to us to have one sort of question answered rather than another, that is no reason to think that a theory which answers more of the first sort of questions is more likely to be true (not even with the

[1] Depending on your views about explanation, this argument may need an additional premise: something to the effect of, "if true, hypothesis *h* would provide a satisfactory explanation of $f_1 - f_n$." However, nothing here turns on its inclusion.

[2] This particular list is due to Quine and Ullian, with 'generality' substituted for the more awkward 'fecundity' (Quine and Ullian 1978, 64-82). It is not unusual. For very similar ones, see (Lycan 1988, 130) and (Lipton 2004, 122). Obviously, not everyone characterizes IBE this way. For example, at best, Bayesians construe IBE as a heuristic tool for fixing the priors and likelihoods. The debate with the Bayesian is an important one, but I can set it aside here: I am taking for granted the conception of IBE that my interlocutors are taking for granted (at least with respect to the objection that I discuss in the main text).

> proviso 'everything else being equal'). It is merely a reason to prefer that theory in another respect (Van Fraassen 1980, 87).[3]

IBE faces a slew of objections, many of which are formidable. However, let's bracket all but the one that appears above. Does it provide a *distinct* problem for those who regard IBE as a source of epistemic justification? In other words, is IBE faulty simply because it relies on "specifically human concerns" that are "a function of our interests and pleasures?" No.

Let's get clearer about van Fraassen's argument. It seems to go as follows:

P1: If reason *r* is an epistemic reason for (subject *s* to believe) *p*, then *r* increases the likelihood of *p*'s truth.
P2: But IBE's reason(s) for (subject *s* to believe) hypothesis *h* do not increase the likelihood of *h*'s truth.
C: Therefore, the reasons given by inference to the best explanation are not epistemic reasons.

Let's call this *the argument from the truth-conduciveness of epistemic reasons* (ATER). If ATER is sound – and if (plausibly enough) you need epistemic reasons to get epistemic justification – then it seems that IBE cannot provide us with epistemic justification.

Some respond to ATER by attacking P2.[4] However, I tend to think that P2 is true. My quarrel is with P1: it is *not* the case that all epistemic reasons increase the likelihood of truth.[5] The claim that they do assumes a form of epistemic value monism – which, I'll argue, even IBE's critics should reject. In my view, then, the objection above amounts to the observation that

[3] We find the same argument in one of Scott Shalkowski's recent papers: "Reasons are sometimes epistemic, sometimes pragmatic. IBE is proposed as a general kind of inference involving epistemic reasons; it is to provide us with reasons to adopt a theory as more likely to be true than its competitors and not merely as a tool useful for accomplishing some non-alethic goal. [...] Simplicity is a theoretical virtue, let us grant, but it is an instrumental virtue. Simple theories are easier to work with, so recognizing that a theory is simple provides one with a reason to work with the theory, but this is a conclusion of a piece of practical reasoning" (Shalkowski 2010, 171-172).

[4] Richard Swinburne, for example, contends that "it is a fundamental a priori principle" that simpler theories are more likely to be true than are more complex ones (Swinburne 2001, 102). And there are a number of less radical defenses of simplicity: e.g., (Quine and Ullian 1978), (Sober 1981), and (Kelly 2007).

[5] There are, of course, philosophers who insist that the individual virtues are individually truth-conducive. Simplicity is usually taken to be the hardest one to defend, but Richard Swinburne nevertheless contends that "it is a fundamental *a priori* principle" that simpler theories are more likely to be true than are more complex ones (Swinburne 2001, 102). There are also a number of less radical defenses of simplicity – e.g., (Quine and Ullian 1978, 69-70), (Sober 1981, 145), and (Kelly 2007, 561).

IBE is incompatible with epistemic value monism, and that is no objection at all.

2. From ATER to TRUE

We need to begin by distinguishing two ways of interpreting ATER's first premise:

> P1: If reason *r* is an epistemic reason for (subject *s* to believe) *p*, then *r* increases the likelihood of *p*'s truth.

On the flat-footed reading, P1 imposes a necessary *external* condition on epistemic reasons: namely, that they must increase the likelihood of truth. But if this is the correct reading, then there are two strikes against van Fraassen's relying on P1. First, no proponent of IBE needs each explanatory virtue to be *individually* truth-conducive; IBE does not require that, for example, simpler theories are more likely to be true just in virtue of their simplicity. Rather, the proponent of IBE needs it to be the case that the virtues are *jointly* truth-conducive. There is no obvious reason why various non-truth-conducive virtues might not 'cancel one another out', allowing the reasoner to triangulate the truth, as it were.[6] Second, and more importantly, while epistemic reasons may need to satisfy an external condition, it's hard to see how making this point would help van Fraassen. To assess IBE's reliability, we would need to check whether (a) there is a positive correlation between those propositions supported by the reason in question and those propositions that are true and (b) a negative correlation between those propositions supported by the reason in question and those propositions that are false. But propositions don't wear their truth values on their sleeves, so we can only go on our best judgments. And as soon as we admit this, we must also recognize that there will be disagreement: if I think that our judgments about the existence and properties of unobservables are generally accurate, then I will be inclined to say that reasons supporting those judgments are truth-conducive; if van Fraassen doesn't, then he

[6] For more on this point, see (Fischer, ms).

won't be so inclined. So on the 'necessary external condition' reading of P1, the merits of ATER turn on the merits of P2 (which alleges that IBE's reason(s) for hypothesis *h* do not increase the likelihood of h's truth). And our judgments about the merits of P2 will depend on two factors: first, the list of hypotheses that we believe to be justifiable via IBE; and second, the list of those hypotheses that we take to be true. But van Fraassen is using this argument (among many others) to *motivate* shortening the list of hypotheses that we take to be true; i.e., he is trying to argue that IBE cannot justify beliefs about unobservables. Hence, the 'necessary external condition' reading of P1 does not help his project; it seems to beg the question at hand. What is the alternative?

I propose that P1 concerns the *aim* of epistemic reasons. We might reformulate ATER accordingly:

P1*: If reason *r* is an epistemic reason for (subject *s* to believe) *p*, then *r* is aimed at increasing the likelihood of *p*'s truth.
P2*: But IBE's reason(s) for (subject *s* to believe) hypothesis *h* are not aimed at increasing the likelihood of *h*'s truth.
C: Therefore, the reasons given by inference to the best explanation are not epistemic reasons.

I think that this reading fits more naturally with the passage quoted above; at any rate, it avoids the problems just mentioned. It also make it clear why, earlier, I posited a connection between P1 and epistemic value monism. P1* insists that the only epistemically valuable feature of a reason is its being aimed at truth. Hence, P1* commits its proponent to a version of epistemic value monism – the view that reasons have only one epistemically relevant feature.

3. TRUE

What's wrong with P1*? It will be easier to see this if we take a detour through ethical theory.[7]

[7] My argument in this and the next section is inspired by (Lycan 1988, Chapter 7). I do not mean to suggest that

Utilitarianism, at least in its simple, hedonistic form, is committed to both *value monism* and *proceduralism*. An ethical theory is committed to the former if it maintains that all situations have only one morally relevant feature; according to utilitarianism, that feature is well-being. An ethical theory is committed to the latter just in case it says that there is a decision procedure for determining whether an action is obligatory, permissible, or wrong; for utilitarians, this is the principle that you should maximize well-being.[8] With these two points in mind, and idealizing a bit, we can represent utilitarianism with a function: it takes a set of action / outcome pairs as inputs, selects the one with the greatest overall well-being, and gives the action that leads to that situation as the output; that action, of course, is the one that utilitarians judge to be obligatory.[9]

The function just outlined represents act utilitarianism. How would we need to modify it in order to represent rule utilitarianism? We replace the set of action / outcome pairs with a set of slightly more complex pairs, the first member of which is a candidate moral rule, the second of which is the outcome that would result from universal adherence to that rule. The function still selects the one with the greatest well-being. However, instead of giving an obligatory action as an output, it gives a moral rule; we then apply the rule to our situation to determine what's obligatory.

It's easy to reframe rule utilitarianism as an epistemological position. Instead of candidate moral rules, the first member of each pair is a candidate epistemic policy; instead of global outcomes, the second member of each pair is the number of truths that would be believed if that policy were followed.[10] Instead of selecting the outcome with the greatest well-being, our

---

Lycan would agree with anything that I say here.

[8] I'm using the phrase 'decision procedure' loosely, where it doesn't imply that informed and competent agents are always in a position to carry it out.

[9] Here is one respect in which this is an idealized representation: like Stalnaker's account of counterfactuals, it assumes that there are no ties.

[10] I am treating wellbeing as a simple property; hence, the parallel with truths believed. Later, I'll discuss a variant that balances truths believed with falsehoods avoided.

new function selects the one with the greatest number of resultant true beliefs, giving that epistemic policy as an output. As before, the output is not itself the obligatory action; rather, the output is the principle that determines what you ought to believe in a given circumstance – equivalently (so say I), it determines the belief that you would be justified in holding in those circumstances. Let's call this view *truth-maximizing rule utilitarianism in epistemology* (TRUE). Like its cousin, TRUE is a version of value monism: it takes truth to be the only feature of a belief that is of worth. Also like its cousin, TRUE is a form of proceduralism: it takes there to be a straightforward decision procedure that settles which of the many possible epistemic policies is correct. Why does it recommend maximizing true beliefs? As in ethics, your theory of value drives your theory of the right: if you think that only well-being is of moral worth, then it is hard to see what you would recommend *other* than maximizing well-being. After all, if well-being is of moral value, then surely more is better, at least if all other things are equal. And if value monism is true, then all other things always *are* equal – there is never anything else with which well-being competes. So, you should maximize it. The same argument applies, *mutatis mutandis*, to truth given TRUE.

TRUE is probably not just a form of epistemic value monism: it is probably the only epistemological position that is plausible if epistemic value monism is correct. As I suggested in the preceding paragraph, it's likely that epistemic value monists are committed to an epistemology that is structurally analogous to utilitarianism. But in epistemology, the analog of act utilitarianism is hopeless: *that* view would say that a belief is justified iff it's true, since (a) such a view would only take into consideration the local features of the belief and (b) such a view would take the truth of that belief to be the only feature that matters. But, of course, it is not the case that a belief is justified iff it's true. The analog of rule utilitarianism, TRUE, avoids this

problem by introducing the epistemic policies: they are designed to take non-local factors into account – namely, the number of true beliefs that would be achieved given universal adherence to the epistemic policy – thereby preventing TRUE from having the awkward consequence that sinks the epistemic analog of act utilitarianism.

4. We Should Reject TRUE

However, as sane as it may sound, TRUE has very implausible implications. Here is the argument. I suggested that we can represent TRUE as a function: the inputs are policy / success rate pairs, the output is the most truth-conducive epistemic policy. I also intimated that 'being the most truth-conducive epistemic policy' means 'being the policy that produces the greatest number of true beliefs if it were followed'. But this can't be right. The policy that will do best here is the one that tells us to believe *everything*. If truth is the *only* valuable doxastic feature, then there is no value to avoiding falsehood. So, if we were to believe every proposition and its negation, then we wouldn't miss out on any truths, thereby maximizing what's of epistemic value.[11] But this is ridiculous.

To avoid this problem, we should make a friendly amendment to TRUE. We'll still say that truth is still the only valuable doxastic feature, but we'll add a principle called 'NOFALSITY', according to which believing falsely has epistemic *dis*value. Call our revised version of TRUE – i.e., the conjunction of TRUE and NOFALSITY – 'T&~F'. T&~F preserves the spirit of TRUE, if not the letter. Problem solved?

No. Now, the most straightforward interpretation of 'being the most truth-conducive

[11] *Objection:* We can't believe contradictions, epistemic policies create epistemic obligations, and we aren't obligated to do the impossible; so, we can't be obligated to believe every proposition and its negation, which means that this policy is not in the running. *Reply:* It's not at all clear to me that we can't believe explicit contradictions. But even if that's right, then we certainly *can* believe implicit contradictions. In other words, even if we can't believe p & $\sim p$, it's surely the case that we can believe p and we can believe $\sim p$.

epistemic policy' is something like 'being the policy that maximizes the ratio of truths to falsehoods believed'. This looks like a recipe for radical epistemic caution: if you take this policy seriously, then you should believe only self-evident truths. If you believe even one falsehood, then it doesn't matter how many truths you believe, since your ratio of true to false beliefs will invariably be lower than it would have been had you believed no falsehoods at all. But as long as you find at least one self-evident truth (the *cogito* or your favorite tautology) and you believe no falsehoods whatever, your ratio will be as high as it possibly can be.[12]

So we seem to be torn between two extremes: either radical epistemic abandon (believe everything) or radical epistemic caution (believe only the self-evident). You might object I'm assuming both more and less control over our beliefs than is plausible, ignoring:

(a) that you can't believe whatever you want, so you can't believe everything (which is supposed to undermine my objection to TRUE), and

(b) that so many of our beliefs form spontaneously, so we can't limit ourselves to believing a single self-evident truth (which is supposed to undermine my objection to T&~F).

I grant both (a) and (b), but they make no trouble for my argument. Concerning (a), is it really just your *inability* to believe everything that makes it a terrible epistemic policy? If TRUE is correct, then this seems to be the case. Surely it isn't, though. Even if it *were* psychologically possible to believe indiscriminately, that would not be a way of securing justified beliefs. And the same point applies to (b): even if it *were* psychologically possible to believe only the self-

[12] *Objection:* Any number over zero isn't a ratio (it's ill-defined); so, you would have to believe at least one falsehood to achieve the goal of maximizing the *ratio* of true to false beliefs. *Reply:* First, the 'maximize the ratio' formulation isn't mine; it's common enough in the literature: see, e.g., (Nozick 1993, 69). Second, it's easy enough to recast the conversation in terms of maximizing the percentage of truths believed, in which case my argument stands. And third, you can still make the ratio version work. Suppose that you believe one self-evident truth and believe its negation; you then believe as many propositions as you can that are logically equivalent to the self-evident truth. Since there are infinitely many of them, you can make the ratio as high as your mental capacities permit (and this with minimal epistemic risk). *Objection:* Logically equivalent propositions are equivalent, period; so, this solution puts your ratio at .5. *Reply:* Logically equivalent propositions are not equivalent, period. If they were, then 'red is a color' and '2 + 2 = 4' would express the same proposition, since they both express necessary truths. And that's absurd.

evident, would this be a good epistemic policy? If T&~F is correct, then the answer is 'Yes'. But surely this would be a mistake.

Here is a further consideration. Perhaps some beliefs are inescapable: even if we judge them to be false, we cannot abandon them. If there are such beliefs, though, and we indeed judge them to be false, then surely we can still recognize the epistemic tension that this creates. I suspect that something similar is the case when we judge the risk of error to be unacceptably high: whether or not we can actually abandon the beliefs in question, if we judge the risk of error to be too great, then surely we can judge them to be epistemically subpar. But when is the risk excessive? If TRUE is correct, then our only advice is to believe as many truths as possible; it follows that the risk is *never* excessive. If a belief is epistemically subpar just in case the risk of being wrong crosses some threshold, TRUE seems to suggest that we should *never* judge a belief to be epistemically subpar. Alternately, if T&~F is correct, then our only advice is to maximize the ratio of truths to falsehoods believed; now, the risk is excessive whenever there is a threat that we might *not* maximize that ratio, which is to say that it's excessive whenever we believe what isn't self-evident. T&~F seems to suggest, then, that we should almost *always* judge our beliefs to be epistemically subpar. So, whether supplemented with NOFALSITY or not, TRUE is in trouble.

Someone might object that it's uncharitable to articulate either TRUE or T&~F in terms of truth *simpliciter*. Rather, they should be cashed out in terms of *significant* truths (i.e., "maximize the number of significant truths believed' or 'maximize the ratio of significant truths to falsehoods believed'). I agree that it should be, but the proponents of TRUE and T&~F cannot. What makes some truths significant while others are not? Whatever it is, it's something other than their mere truth – perhaps their usefulness, or their explanatory power, or their fit with

what we believed pre-theoretically, or what have you. And crucially, either the significance of a belief is explicable solely in terms of its truth, or it isn't. In other words, significance is either extrinsically or intrinsically valuable. If it's extrinsically valuable, then significance won't save either TRUE or T&~F from the problems that I've been detailing, since there will never be a case in which significance trumps truth, thereby giving you a reason to take an epistemic risk. But if significance *isn't* explicable solely in terms of its truth – i.e., if it's intrinsically valuable – then to set significance alongside truth is to reject epistemic value monism, and hence to reject TRUE and T&~F.

5. Back to IBE

I grant that I may have overlooked a perfectly good policy that's based on the assumption that truth is the only thing of epistemic worth; if so, then TRUE's devotees should provide it. Suppose they can't. How should we diagnose the problem? Well, as I've indicated, rationally increasing your stock of beliefs beyond the self-evident requires a policy about the management of epistemic risk. Whatever policy you adopt, it will need to give advice having the following form: *risk error only if…*, where the ellipses stand for something *else* of epistemic worth. Your policy might be, for example, that you should risk error only if the proposition would increase the coherence of your belief system. Alternately, you may maintain that you should risk error only if the proposition in question seems to be true, absent any defeaters – this is the way taken by those who defend 'phenomenal conservatism' (e.g., (Huemer 2001)). If you go the first route, then you're assigning epistemic value to coherence; if you go the second, then you're assigning it to conservatism. There are no doubt plenty of other options, but they'll all lead you to assign intrinsic epistemic value to something other than truth. In other words, they'll lead you to deny

epistemic value monism. But now recall ATER:

P1*: If reason *r* is an epistemic reason for (subject *s* to believe) *p*, then *r* is aimed at increasing the likelihood of *p*'s truth.
P2*: But IBE's reason(s) for (subject *s* to believe) hypothesis *h* are not aimed at increasing the likelihood of *h*'s truth.
C: Therefore, the reasons given by inference to the best explanation are not epistemic reasons.

If epistemic value monism is false, then P1* is false. So P1* is false. IBE may face a number of serious challenges, but ATER is not among them.

References

Kelly, Kevin. 2007. A New Solution to the Puzzle of Simplicity. *Philosophy of Science* 74:561-73.

Lipton, Peter. 2004. *Inference to the Best Explanation.* International Library of Philosophy. New York: Routledge.

Lycan, William G. 1988. *Judgement and Justification.* New York: Cambridge University Press.

Quine, W. V. and J. S. Ullian. 1978. *The Web of Belief.* New York: Random House.

Shalkowski, Scott. 2010. IBE, GMR, and Metaphysical Projects. Pages 167-187 in *Modality : Metaphysics, Logic, and Epistemology*. Edited by Bob Hale and Aviv Hoffmann. New York: Oxford University Press.

Sober, Elliott. 1981. The Principle of Parsimony. *British Journal for the Philosophy of Science* 32:145-56.

Swinburne, Richard. 2001. *Epistemic Justification.* New York: Oxford University Press.

Van Fraassen, Bas C. 1980. *The Scientific Image*. New York: Oxford University Press.

Title: Broken Mechanisms: Function, Pathology, and Natural Selection

Abstract: The following describes one distinct sense of 'mechanism' which is prevalent in biology and biomedicine and which has important epistemic benefits. According to this sense, mechanisms are defined by the *functions* they facilitate. This construal has two important implications. Firstly, mechanisms that facilitate functions are capable of breaking. Secondly, on this construal, there are rigid constraints on the sorts of phenomena 'for which' there can be a mechanism. In this sense, there are no 'mechanisms for' pathology, and natural selection is not a 'mechanism of' evolution, because it does not serve a function.

Section 1. Introduction. The following presents a distinct sense of 'mechanism' that is prevalent in biology and biomedicine and which has important epistemic benefits. I will use the term 'functional mechanism' to describe this sense. According to this sense, a mechanism is defined by the function that it serves (in addition, perhaps, to other characteristic features such as spatial, temporal, organizational, and hierarchical constraints). More formally, for all *X* and for all *Y*, where *X* is a biological system and *Y* is a biological phenomenon, *X* is a mechanism for *Y* only if *X* has the function of facilitating *Y*. Strictly, this is not a definition of 'mechanism' but a necessary condition on the sense of 'mechanism' I wish to identify. This is a sense that has been obscured or overlooked in much of the new mechanism literature, though some biologists, psychologists, and philosophers have recognized it explicitly (Williams 1996, 9; Tooby and Cosmides 2006, 185; Buss 2005, 69; Moghaddam-Taaheri 2011; Moss 2012).

There are two important implications of this characterization of mechanism. These implications can also be used as indicators of its presence in biological contexts. First, mechanisms that serve functions can break. To say that a mechanism for *Y* is 'broken' means that *Y* is its function and it fails to perform *Y*. Moreover, it is difficult to understand what else it might mean for a mechanism for *Y* to be 'broken,' rather than for it to cease to be a mechanism for *Y*. Biologists and biomedical researchers have a comprehensive lexicon to describe ways that mechanisms can break. A mechanism can 'breakdown;' it can be 'usurped' or 'coopted' by another mechanism or biological process; it can be 'interfered with' or 'disabled;' it can 'fail to function.' Philosophers of the new mechanism tradition have recognized the fact that mechanisms can break, and

have described its significance for understanding causation, identifying the components of mechanisms, and treating disease (e.g., Bechtel and Richardson 1993, 19; Craver 2001, 72; Glennan 2005, 448; Darden 2006, 259). Consequently, it is imperative to understand the commitments this involves.

Second, functional mechanisms impose constraints on the sorts of biological phenomena 'for which' there can be a mechanism. For example, in this sense of the term, there are no 'mechanisms for' pathology, because a pathology in a biological system is not a function of any part of that biological system.[1] Rather, pathologies are explicable as causal consequences of the *breakdown* of a mechanism for a function. Secondly, in this sense, natural selection is not a 'mechanism' of evolution because it does not serve a function, on any well-developed theory of function that is consistent with biological usage. This statement will be defended in Section 4.

The following adopts a modest pluralism with regard to 'mechanism.' There are cases in which biologists and biomedical researchers use the term 'mechanism' without functional implications. In some cases, mechanism is used synonymously with 'physical explanation' (Moss 2012); in this sense of the term, it is almost trivial to say that natural selection is a 'mechanism of' evolution, just as it is a 'mechanism of' extinction. My concern, however, is that the functional sense of mechanism has been obscured by much

1 It may be the case that there is a mechanism in some other system that performs its function by inducing a breakdown in a mechanism in the first system, which in turn causes a pathology. (I thank Joyce Havstad for this observation.)

of the new mechanism literature. Glennan, for example, has insisted that 'mechanism' has no normative or teleological connotations (Glennan 1996, 52-3; 2002, 128; 2005, 445). Craver accepts that mechanisms serve functions but accepts an extremely liberal conception of function according to which the function of a system is relative to the interests of the research community that investigates it (see, also, MDC 2000, 6; Craver 2001; forthcoming; Glennan 2002, 127 [fn. 6]; Glennan 2005, 456).[2] Both of these commitments are inconsistent with the strain of biological and biomedical usage I wish to identify. More importantly, these commitments tend to misinterpret such usage where it occurs, and by doing so, they relinquish the epistemic benefits associated with this usage.

The following also adopts a modest pluralism with respect to 'function.' My intention is not to identify a uniquely correct sense of 'function.' Rather, there are several concepts of function that are consistent with this sense of 'mechanism,' such as those that appeal to selection, design (as in the case of artefacts), or contributions to the survival, reproduction, or inclusive fitness of individuals.[3] However, it is important to note that this sense of mechanism is not consistent with the causal role theory of function or its more

2 Also see Bechtel and Richardson 1993, 17, where the 'function' of a part is characterized in terms of its causal role – that is, its contribution, in tandem with the other parts, to the 'behavior' of the system as a whole. Glennan (2005; 448) uses the term 'causal role' to characterize the 'function' of a part.

3 See Garson 2011; 2012, which defend a generalized account of the selected effects theory.

recent variants, according to which the function of a system's part consists merely in its contribution, in tandem with the other parts of the system, to some phenomenon of interest to a research community. This is because the causal role theory licenses ascriptions of function, and hence, ascriptions of mechanism, that are inconsistent with much of the biological and biomedical literature, as will be shown in Sections 2 and 3.

The view that mechanisms serve functions is not novel; G. C. Williams (1966) forcefully propounded it in his famous *Adaptation and Natural Selection*. He proposed that 'mechanism' be defined in terms of function; in addition, he held that 'functions' are selected effects (ibid., 9). He maintained that 'mechanism' should not be used to describe incidental effects of a trait or physically inevitable consequences of a trait's performing its function (ibid., 11-12). The reason for his insistence is that he regarded the term 'mechanism' as synonymous with 'means,' but the latter concept is inapplicable in the absence of a corresponding function, goal, or purpose. Some evolutionary psychologists have accepted Williams' strictures on the term 'mechanism' (e.g., Tooby and Cosmides 2006, 185; Buss 2005, 69), and similar views have recently been proposed by Moghaddam-Taaheri (2011) and Moss (2012).[4] As will be elaborated in Section 2,

4 The difference between Moss' view and my own is that, according to Moss, mechanisms need not *serve* functions. They need only '*refer* to' the functions (or goals) of a biological system (pers. comm.) For example, in my view, there are no mechanisms for pathology because pathologies on the part of a system are not a function of that biological system. On Moss' view, there are mechanisms for pathology because to describe something as a 'mechanism for pathology' is to make *reference* to the goals of a biological system (since, by definition, pathologies tend to undermine the ability of such

Williams' usage is consistent with much contemporary biological usage as well. As a consequence, the working assumption that mechanisms serve functions is a useful heuristic for philosophers, sociologists and historians to employ in the interpretation of biological texts.

Section 2 describes the prevalence of this sense of 'mechanism' in biology and biomedicine and its epistemic benefits. Sections 3 and 4 respond to two kinds of counterexamples that purport to show that this view is largely at odds with biological usage.

Section 2. Functional Mechanisms. This section does three things. First, it shows how the notion of a functional mechanism provides a parsimonious explanation for how mechanisms can break. Second, it shows the prevalence of this sense of mechanism in biological and biomedical usage. Finally, it describes an important epistemic benefit of this usage, namely, that it maximizes the inferential coherence of biology and biomedicine.

Conceptual Parsimony

The fact that (some) mechanisms can break implies that (some) mechanisms are normative. To say that a mechanism is 'normative' simply means that, where *Y* represents some biological phenomenon of interest to researchers, it is possible for something to be

systems to realize their goals).

a mechanism for *Y*, despite the fact that it cannot perform *Y*. One way to explain the normativity of mechanism is by reference to the normativity of function. That is, to say that a mechanism for *Y* is 'broken' implies that has the function of facilitating *Y* but cannot do so.

Moreover, this is a particularly *good* explanation for the normativity of mechanism because it exhibits conceptual parsimony. Although the proper explication of 'function' is controversial, there is no great mystery about how functions can be normative (see Garson 2008 for an overview). For example, according to the selected effects theory, to say that biological trait *X* has the function *Y* is to imply that *X* was selected for *Y* by a natural process of selection. To say that *X* is dysfunctional with respect to *Y* implies, amongst other things, that it cannot do *Y*. For another example, according to one version of the 'goal-contribution' account, the function of a trait is defined, roughly, as its statistically typical contribution to the survival or reproductive capacity of each member of the reference class that possesses that trait. To say that a trait cannot perform its function implies that it cannot make this contribution.

As a consequence, philosophers have availed themselves of the concept of function in explaining the normativity of other biological categories, such as biological 'information' and biological trait classification. For example, for some philosophers, to say that a signal can carry 'misinformation' implies that it fails to fulfill its proper function of indicating the source (e.g., Dretske 1986). For another example, to say that biological trait classification is 'abnormality-inclusive' is to say that what makes a token of a trait a

member of a certain type is the fact that it possesses the function that defines the type, even if it is unable to perform that function (e.g., Neander 1991; Rosenberg and Neander 2009). This observation is not necessarily an endorsement of these approaches to information and trait classification. The fact that appeals to the concept of function are plausible and defensible in other biological contexts suggests, however, that it is a parsimonious strategy for explaining the normativity of mechanism. Below, I will explain how this strategy is also quite useful in biomedicine.

This is not to say that biologists do not sometimes use the term 'mechanism' in some other sense, one without normative connotations. However, to the extent that it makes sense to talk about a 'broken' mechanism, it is likely that the functional sense of 'mechanism' is presupposed.

Consistency with Biological Usage

Biologists routinely explain pathologies by reference to the 'breakdown,' 'cooption,' 'usurpation,' or 'interference with,' a biological mechanism. A short list of examples can be used to illustrate the point:

- "…drugs of abuse can *hijack* synaptic plasticity mechanisms in key brain circuits." (Kauer and Malenka 2007, 844; emphasis mine)

- "…drugs of abuse can *co-opt* synaptic plasticity mechanisms in brain circuits involved in reinforcement and reward processing." (ibid.; emphasis mine)

- "Only by understanding these core synaptic mechanisms can we hope to understand how drugs of abuse *usurp* or *modify* them." (ibid., 845; emphasis mine)

- "It is argued here that potentially irreversible *impairments* of synaptic memory mechanisms in these brain regions are likely to precede neurodegenerative changes that are characteristic of clinical [Alzheimer's disease]." (Rowan et al. 2003, 821; emphasis mine)

- "However, it is possible that a *disruption* of synaptic plasticity-related mechanisms by soluble AB also contributes to clinical symptoms." (Ibid., 826; emphasis mine)

The fact that biologists often explain diseases in terms of broken mechanisms suggests that, for these cases, mechanisms are defined by the functions they serve. It does not imply that biologists *always* define mechanism in terms of function. But the fact that they sometimes do so demands an explanation. The explanation offered here is that they are utilizing the functional sense of mechanism.

Craver (2001; forthcoming) develops a view of the relationship between mechanism and function according to which mechanisms serve functions (also see Piccinini and Craver 2011 and MDC 2000, 6). His attempt to give an explicit and lucid account of the relation between mechanism and function is admirable. However, his particular construal of the concept of function is overly liberal. I believe that Craver's overly liberal concept of function has the tendency to distort biological usage in important ways, and in so doing, to forgo the epistemic benefits of this more restrictive use. Craver accepts a version of the causal role theory associated with Cummins (1975). For Craver, all that is required for an activity of a system (considered *in toto*) to constitute its function is for there to be a research community which takes that activity as the focus of its explanatory interest. Once the research community has (conventionally) selected an activity of the system to constitute its function, the function of each *part* of the system can be identified (non-conventionally) as the contribution that it makes, in tandem with the other parts, to yielding the function of the system as a whole.

However, this 'perspectival' view of function is inconsistent with much of biological and biomedical usage. For example, if a research community is interested in the pathophysiology of Alzheimer's disease, then, according to Craver, it would be appropriate, from the standpoint of that research community, to say that certain neurological processes have the function of producing Alzheimer's disease, and that the mechanisms that carry out this function are 'mechanisms for' Alzheimer's disease (Craver forthcoming). But Alzheimer's disease is almost universally recognized as a 'dysfunction,' and the causal processes that produce it are often described as the result of

a broken 'mechanism for' normal cognitive function, as indicated in the quotations above. I do not claim that Craver cannot develop his theory in such a way as to make sense of this discrepancy (see, e.g., Hardcastle 1999 for such an attempt, though I believe that Hardcastle's attempt also results in function ascriptions that are inconsistent with biological usage). My claim, however, is that there is no correlation between the fact that a research community takes an interest in a phenomenon and the willingness on the part of the members of that community to describe that phenomenon as a 'function' of some system and the causal processes that carry out that function a 'mechanism for' that phenomenon.

Epistemic Benefits

Lastly, and most importantly, the notion of a functional mechanism has epistemic benefits. It is a good habit of thought for biologists and biomedical researchers. This is because it maximizes the inferential coherence of biology and biomedicine. In short, there are many more states of an organ or organ system that are consistent with pathology than are consistent with normal functioning. Moreover, these pathological states typically *can* be explained as the result of broken mechanisms for normal function. This suggests that an efficient research strategy for pathology is to attempt to understand the (relatively) smaller number of mechanisms for normal function and to use that information to both explain the diversity of pathological states (e.g., Moghaddam-Taaheri 2011, 608-610) as

well as predict the existence of pathologies that may not have been discovered or the etiology of which is unknown. [5]

For example, Lambert-Eaton syndrome and myasthenia gravis are two pathologies of the neuromuscular junction. The former impairs the motor neuron's ability to release acetylcholine (ACh) and the latter impairs the muscle fiber's ability to respond to ACh. A researcher may track the etiology of each disease by describing a separate 'mechanism' for each, replete with spatial, temporal, organizational, and hierarchical constraints. Alternately, he or she may track the etiology of each disease by noting that both result, in an explicable way, from breakdowns in the mechanism for ACh transmission in the neuromuscular junction. The latter is more useful because it forces the researcher to integrate information regarding each disease with information about how the mechanism normally functions, in such a way that information about the former enhances information about the latter, and vice versa. This is what I mean by maximizing the inferential coherence of biology and medicine. By the same token, many diseases, such as anencephaly, spina bifida, and cranioachischisis, result from various breakdowns in the mechanism for neurulation. Attempting to identify a separate 'mechanism' for each (again, replete with spatial, temporal, organizational, and hierarchical constraints) is less

5 This point is also suggested in Neander (forthcoming), who argues that the practice of pathology is best served by characterizing pathologies as involving deviations from normal function. While she does not specifically discuss mechanism, I believe the same point can be made with regard to mechanism: pathologies are most efficiently described as resulting from breakdowns in functional mechanisms.

efficient than observing that all of them are explicable consequences of a breakdown in the same mechanism, seeking to identify that mechanism, and identifying the causal pathways by which breakdowns in that mechanism lead to disease. The working assumption that mechanisms are defined by the functions they facilitate helps to standardize that practice.

I am not claiming that all pathologies can be explained currently in terms of broken mechanisms. This is because the mechanisms may be unknown, or the functions of those mechanisms may be unknown. For example, prion-related diseases were believed to be caused by proteins before it was known what mechanism or mechanisms they disrupt.[6] (As it turned out, the prion coopts the folding pattern of other proteins, which disrupts the ability of the latter to carry out their functions.) In these cases, it should be acknowledged that the pathology is likely the result of the breakdown of an *unknown* mechanism, or of the breakdown of a mechanism for an unknown function, rather than that there is a 'mechanism for' the disease.

I am also not claiming that knowledge of the functional mechanism is a logical or epistemological prerequisite for medical treatment. Rather, when such knowledge is available, understanding pathology in terms of broken mechanisms enhances the inferential coherence of the theoretical infrastructure of biomedicine, which may result in improved treatment for that or related pathologies. Finally, I am not claiming that one cannot use information about pathology to illuminate the mechanism for normal function.

6 I thank Lindley Darden for this observation.

Certainly, one can use what is known about the pathology (for example, that cystic fibrosis is associated with mutations in the *cftr* gene) to assist in discovering the mechanisms for normal function. Once these corresponding mechanisms are known, the foregoing considerations suggest that viewing pathology in terms of broken mechanisms is theoretically and practically advantageous.

Section 3. Apparent Counterexample 1: Mechanisms for Pathology. One main criticism of this view is that there are many counterexamples to this usage. Though biologists *sometimes* describe disease in terms of a 'breakdown' of a mechanism, they often use the expression 'mechanism for pathology.' This fact can be confirmed by doing a search for 'mechanism for' in any major biological or biomedical journal. This suggests that, as a rule, the use of the term 'mechanism' is independent of considerations of function.

It is true that there are numerous *apparent* counterexamples to this view, that is, instances in which biologists use the locution, 'mechanism for pathology.' However, this expression can often be seen, justifiably, as elliptical for one that has a different signification than that which philosophers of the new mechanism tradition would generally attribute to it. Specifically, when a biologist claims to have discovered a 'mechanism for pathology,' that locution can often be interpreted as shorthand for the claim that there is a *mechanistic explanation* for the pathology. The term 'mechanistic explanation' is used here non-conventionally to describe an explanation that cites a mechanism. As argued above, pathologies typically do admit of 'mechanistic explanation'

in this non-conventional sense, because they can often be explained via a breakdown or cooptation of a mechanism and hence by reference to a mechanism. But in this sense, to say that there is a 'mechanistic explanation for' *Y* does not imply that there is a 'mechanism for' *Y*. All it implies is that there is a mechanism for some function *Z*, and *Y* results from the breakdown of this mechanism.

For example, two recent popular presentations of scientific articles *seem* to recognize the existence of mechanisms for pathologies.[7] The first is entitled, "Team Identifies Mechanism of Cancer-Induced Bone Destruction;"[8] the second, "A Possible Physical Mechanism of Cancer Metastasis."[9] However, a careful reading of the articles on which they are based shows that they actually support the view that mechanisms serve functions. For example, in the scientific article on cancer metastasis, the mechanism identified, *and described as a 'mechanism,'* is merely a mechanism for cell elasticity. This property has functional significance but can be coopted in such a way as to facilitate metastasis (Rolli et al. 2010). In the article on bone destruction, the mechanism described, *and described as a 'mechanism,'* is a mechanism for bone resorption, which along with bone formation performs the function of maintaining bone structure (Lynch et

7 I thank Stuart Newman for these references.

8 http://www.mc.vanderbilt.edu/reporter/index.html?ID=3979, accessed October 29, 2011.

9 http://www.nasw.org/users/mslong/2010/2010_01/Metastasis.htm, accessed October 29, 2011.

al. 2005). It explains bone destruction in terms of the dysregulation of the balance between formation and resorption.

These articles suggest that when biologists talk of a ‘mechanism for’ pathology, the mechanism in question should often be understood not as a ‘mechanism for’ the pathology but a ‘mechanism for’ a lower-level component within a pathological system, which when considered on its own may have functional significance but which may be coopted to produce pathology. This form of explanation can loosely be called a ‘mechanistic explanation’ because it cites a mechanism. One need not recognize mechanisms for pathology in order to accommodate this usage.

Despite the fact that many apparent counterexamples are not actual counterexamples, actual counterexamples to this view probably exist. However, the existence of *actual* counterexamples should not be taken to discredit the theory as a whole, so long as these counterexamples are infrequent. This is because the property of facilitating a function is not a necessary condition for characterizing *every* instance of the term ‘mechanism’ in the biological literature, but only a distinctive and prevalent subset. Along the same lines, some of the founding documents of the new mechanism tradition emphasize that the various definitions of ‘mechanism’ offered are not intended as necessary and sufficient conditions for use, but as characterizations that emerge from philosophical reflection on biological usage (e.g., Darden 2006, 273). The proposal offered here should be taken in a similar spirit. One consequence is that the mere existence of isolated counterexamples need not disqualify this proposal; in the same way, the fact that scientists do not always

use the term ‘mechanism’ with the rich spatial, temporal, organizational, and contextual constraints associated with the new mechanism tradition need not disqualify the latter.

Explications of biological ‘mechanism’ should be judged, not in terms of their consistency with every conceivable instance in which a scientist uses the term, but in terms of the benefits and costs of accepting the proposed usage. The last section presented three benefits associated with the notion of a functional mechanism. The cost is that there may be occasional counterexamples that cannot be accommodated in the prescribed fashion. In order to discard this view, one would at least have to show that the actual (and not merely apparent) counterexamples are numerous enough to render the analysis largely inconsistent with biological usage, to the point where the benefits are outweighed by the fact that it often produces misunderstandings, that it thwarts philosophical attempts to understand the way biologists reason about the world, or that it does not constitute a good methodological strategy for biology.

Section 4. Apparent Counterexample 2: Natural Selection as a ‘Mechanism’.

Scientists often describe natural selection as a ‘mechanism’ of evolution (e.g., Havstad 2011, for examples). This has produced a debate amongst philosophers of biology about whether natural selection is a mechanism in the sense characterized by the new mechanism tradition. Some have argued that it is not a ‘mechanism’ in that sense because it does not exhibit a unique decomposition into parts (e.g., Skipper and Millstein 2005, 336); there is too much variability in the spatial or temporal organization of the parts (Ibid., 338; Havstad 2011); it fails to exhibit the kinds of activities or interactions

characteristic of mechanisms (Skipper and Millstein, 2005, 341); or the stages of natural selection are connected by probabilistic and non-deterministic links (Ibid., 343; also see Darden 2006, 278-9 and Barros 2008 for a response).

According to the sense of 'mechanism' sketched above, natural selection is not a mechanism of evolution because natural selection does not have a function. This is the case whether one appeals to the selected effects theory of function or the goal-contribution theory. On the selected effects theory, something has a function only if it was selected for by a selection process. Natural selection itself, however, is not selected for. On the goal-contribution theory, the function of a trait consists in its (statistically typical) contribution to the goal of a biological system in which it is contained (that is, of which it is a component). Though natural selection can promote the evolution of such goal-directed biological systems, it is not in any obvious sense a 'component' within a biological system. There may be another sense of the term according to which natural selection is a 'mechanism,' such as the causal role view, but as noted above, this sense is largely inconsistent with biological usage. The fact that there is a serious disagreement regarding whether or not natural selection is a 'mechanism' of evolution suggests that some of the disputants may have something like functional mechanisms in mind.

While natural selection is not a mechanism, this would not prevent other evolutionary processes from having 'mechanistic explanations' in the functional sense. Mutations, for example, often result from breakdowns of mechanisms for replication, proofreading, or mismatch repair. Williams (1966, 125) shows admirable consistency in his use of

'mechanism' and 'function' when he states that mutation is not a 'mechanism for' producing offspring with new combinations of genes, because mutations do not have a function. It is possible that some mutations result from a functional mechanism (rather than a broken mechanism). There are hypothesized mechanisms for upregulating the mutation rate in the face of environmental stress, via, e.g., the upregulation of error-prone DNA polymerase Pol IV (Darden 2006, 248-267).

The notion of a functional mechanism has important implications both for philosophical discussions about mechanism and for biology and biomedicine. First, it highlights a distinct sense of 'mechanism' that is prevalent in biology and biomedicine and that has been largely neglected. Second, this sense of 'mechanism' maximizes the inferential coherence of biology and biomedicine. Third, it helps to diagnose and resolve various disagreements about the scope of 'mechanism,' specifically, whether there are pathological mechanisms or whether or natural selection is a mechanism.

Acknowledgements: I am grateful to Carl Craver, Lindley Darden, Blaine Ford, Joyce Havstad, Lenny Moss, Karen Neander, Stuart Newman, Anya Plutynski, and Eric Saidel, for comments and criticism of an earlier draft. I am also grateful to the audience members at a session at ISHPSSB 2013, a symposium at the University of Missouri-St. Louis, where some of this material was presented, and the D. C. History and Philosophy of Biology Reading Group where an earlier draft was discussed.

References

Barros, D. B. 2008. “Natural Selection as a Mechanism.” *Philosophy of Science* 75:306-322.

Buss, D. M. 2005. *Evolutionary Psychology: The New Science of the Mind, 3rd ed.* Boston: Pearson.

Craver, C. F. 2001. “Role Functions, Mechanism, and Hierarchy.” *Philosophy of Science* 68:53-74.

Craver, C. F. Forthcoming. “Functions and Mechanisms: A Perspectivalist View.” In *Functions: Selection and Mechanisms*, ed. Philippe Huneman, ??-??. Dordrecht: Synthese.

Cummins, R. 1975. “Functional Analysis.” *Journal of Philosophy* 72:741-65.

Darden, L. 2006. *Reasoning in Biological Discoveries*. Cambridge: Cambridge University Press.

Dretske, F. 1986. “Misrepresentation.” In *Belief: Form, Content, and Function*, ed. R. Bogdan, 17-36. Oxford: Clarendon Press.

Garson, J. 2008. “Function and Teleology.” In *A Companion to the Philosophy of Biology*, ed. S. Sarkar and A. Plutynski, 525-49. Malden, MA: Blackwell.

Garson, J. 2011. “Selected Effects and Causal Role Functions in the Brain: The Case for an Etiological Approach to Neuroscience.” *Biology & Philosophy* 26: 547-565.

Garson, J. 2012. “Function, Selection, and Construction in the Brain.” *Synthese* 189: 451-481.

Glennan, S. 1996. “Mechanisms and the Nature of Causation.” *Erkenntnis* 44:49-71.

Glennan, S. 2002. “Contextual Unanimity and the Units of Selection Problem.” *Philosophy of Science* 69:118-37.

Glennan, S. 2005. “Modeling Mechanisms.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:443-64.

Hardcastle, V. G. 1999. “Understanding Functions.” In *Where Biology Meets Psychology*, ed. V. G. Hardcastle, 27-43. Cambridge, MA: MIT Press.

Havstad, J. C. 2011. “Problems for Natural Selection as a Mechanism.” *Philosophy of Science* 78:512-23.

Kauer, J. A., and R. C. Malenka. 2007. “Synaptic Plasticity and Addiction.” *Nature Reviews Neuroscience* 8:844-58.

Lynch, Conor C., et al. 2005. “MMP-7 Promotes Prostate Cancer-Induced Osteolysis via the Solubilization of RANKL.” *Cancer Cell* 7:485-96.

Machamer, P., L. Darden, and C. F. Craver. 2000. “Thinking about Mechanisms.” *Philosophy of Science* 67:1-25.

Moghaddam-Taaheri, Sara. 2011. “Understanding Pathology in the Context of Physiological Mechanisms: The Practicality of a Broken-Normal View.” *Biology and Philosophy* 26:603-11.

Moss, Lenny. 2012. “Is the Philosophy of Mechanism Philosophy Enough?” *Studies in History and Philosophy of Science Part C* 43: 164-172.

Neander, K. 1991. “Functions as Selected Effects: The Conceptual Analysts’ Defense.” *Philosophy of Science* 58: 168-184.

Neander, K. Forthcoming.

Piccinini, G., and Craver, C. 2011. “Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches.” *Synthese* 183:283-311.

Rolli, Claudio G., et al. 2010. “Impact of Tumor Cell Cytoskeleton Organization on Invasiveness and Migration: A Microchannel-Based Approach.” *PLoS One* 5:e8726.

Rosenberg, A., and Neander, K. 2009. “Are Homologies (Selected Effect or Causal Role) Function Free?” *Philosophy of Science* 76: 307-334.

Rowen, M. J., I. Klyubin, W. K. Cullen, and R. Anwyl. 2003. “Synaptic Plasticity in Animal Models of Early Alzheimer’s Disease.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 358:821-28.

Skipper, R. A., and R. L. Millstein. 2005. “Thinking about Evolutionary Mechanisms: Natural Selection.” *Studies in the History and Philosophy of Biological and Biomedical Sciences* 36:327-47.

Tooby, J., and L. Cosmides. 2006. “Toward Mapping the Evolved Functional Organization of Mind and Brain.” In *Conceptual Issues in Evolutionary Biology, 3rd ed*., ed. E. Sober, 175-195. Cambridge: MIT Press.

Appraising Non-Representational Models

Till Grüne-Yanoff
Royal Institute of Technology, Stockholm
gryne@kth.se

ABSTRACT

Many scientific models are non-representational in that they refer to merely possible processes, background conditions and results. The paper shows how such non-representational models can be appraised, beyond the weak role that they might play as heuristic tools. Using conceptual distinctions from the discussion of how-possibly explanations, six types of models are distinguished by their modal qualities of their background conditions, model processes and model results. For each of these types, an actual model example – drawn from economics, biology, psychology or sociology – is discussed. For each case, contexts and purposes are identified in which the use of such a model offers a genuine opportunity to learn – i.e. justifies changing one's confidence in a hypothesis about the world. These cases then offer novel justifications for modelling practices that fall between the cracks of standard representational accounts of models.

1. Introduction

Philosophers' approaches to appraising models have largely been focused on their representational functions. Models are representations; they are good models to the extent that they are good representations. Various criteria for good representations have been proposed, from isomorphism (van Fraasen 1980) through similarity (Giere 1988) to partial resemblance (Mäki 2009). The implicit assumption underlying these accounts is that models represent real targets – entities or properties that are found in the real world. Without this assumption, none of the assessment criteria for models would have much bite: they require comparing model properties with properties that can be independently observed, measured, or at least indirectly inferred.

This differs notably from the way many modellers describe their own work. Instead of seeking to represent aspects of the real world, they claim to be aiming at constructing possible or parallel worlds that may give relevant insights about the real world in more indirect ways (for an elaboration oft his view, see Sugden 2000). In particular, they claim that these model constructions involve reference to possible processes, possible background conditions, and even possible phenomena or properties. Let me call such models *non-representational models*. Crucially, modellers claim that non-

representational models (at least sometimes) offer a genuine contribution to our knowledge about the real world.

Philosophers, if they treat such cases at all, have by and large appraised such non-representational models as playing merely a heuristic role, for example in "conceptual exploration" (Hausman 1992), "getting acquainted with mechanisms" (Hartmann 1995), "define the extreme of a continuum of cases" (Wimsatt 2007), or facilitating "creative thought" (Holyoak & Thagart 1995). This heuristic justification is weak, because success criteria for such functions are unclear in the extreme. Furthermore, it places the use of non-representational models in the same category as taking a walk, reading the newspaper, or whatever else scientists do in order to inspire themselves to novel theory development. Bunching non-representational modelling together with practices that cannot be rationally accounted for seems an unsatisfactory state, which this paper seeks to repair.

Section 2 offers a characterisation of learning from models, and what kind of hypotheses might be learned from non-representational models. Section 3 employs conceptual distinctions from the discussion of how-possibly explanation, in order to analyse different kinds of possibility claims made with non-representational models. Six kinds of non-representational models will emerge. Section 4 illustrates each kind with a concrete scientific model, and argues that in particular contexts and for specific purposes one learns from each. Section 5 concludes.

2. Learning from Models

Modelling is a set of reasoning practices for cognitively limited beings (Wimsatt 2007). The inferences one can legitimately draw from scientific models are inferences from information already contained in one's set of beliefs.[1] An ideal Bayesian agent would have no use for scientific models. Being very much unlike ideal Bayesian agents, humans often have to rely on models to justify some of their beliefs.

It is in this sense that we can learn from models. Models facilitate their users in making inferences from their own background beliefs. If these inferences affect the model user's beliefs about some other hypothesis, then the model user learned from the model. Learning from a model M, I suggest, is constituted by a *change in confidence in certain hypotheses*, *justified by reference to* M.

[1] Including beliefs one accepts only tentatively, e.g. for the purpose of a thought experiment.

We do not learn from models in the same way as we learn from straightforward observation. Although observation (of the model) is often part of modelling, we ultimately do not want to learn about the model artefact, but about the real world. Thus the learning I will focus on in this paper concerns changes in confidence in a hypothesis *about the world*.

With representational models this is accomplished by (i) investigating certain properties of the model and (ii) establishing that the model is a sufficiently accurate representation of a (real world) target, in order to license an inference from model to target. Aerodynamic behaviour of a scale model of a new type of airplane, for example, is investigated in a wind tunnel. It is then concluded that an actual airplane of that type has similar properties, given that scale model and actual plane are sufficiently similar with respect the proportions of their hull elements, the geometry of their wings, etc. If the model user believes in the truth of the model investigation and the sufficient similarity between model and target, and her prior beliefs about the plane's aerodynamics are not identical to the model result, then she has learned from the model about the world.

I claim that one can similarly learn from non-representational models. That is, reference to non-representational models may justify changing one's confidence in some hypothesis about the world. By definition, this cannot be accomplished by a belief in the model being a sufficiently accurate representation of a (real world) target. Instead, the inference from model to hypothesis must be licensed differently. I will argue that typical beliefs that license such inferences are those that consider certain background conditions or certain processes "possible", or "credible" (Sugden 2000). Hypotheses whose confidence change is justified through reference to such models include the following types:

- That an entity or property is possible. A special case of this is the hypothesis that something is impossible in the actual world.
- That a process yields a property. A special case of this is the hypothesis that an actual process does or does not have the capacity (in non-actual circumstances) to bring about a certain property.
- That an entity or property possibly is a cause of an actual phenomenon

Of course, such hypotheses do not make claims about particular actual entities or about properties instantiated in the real world. To justify changes in such hypotheses would require models that represented these entities or properties sufficiently well. Nevertheless, these hypotheses are about the world. Reflecting on the impact of such hypotheses on explanation or control supports this claim. Consider for example:

- A policy maker seeking to reduce urban segregation might change her policies upon learning that racist preferences are not a necessary cause of segregation.
- A scientists seeking to explain a population dynamic might change his explanatory strategy when learning that this dynamic cannot be produced from actual background conditions with a set of plausible migration decision rules alone.
- A policy maker who learns that preferences for reciprocation are adaptive under certain possible conditions might change her evaluation of certain institutional regulations.

Thus, changes in confidence of hypotheses of the above kind affect the ways we seek to explain and control the actual world. If non-representational models would justify changes in the confidence of such hypotheses, one would learn from such models about the world.

3. How-Possibly Explanations

Schematically, a model consists of a set of initial conditions *Q*, a model process *P* and a model result *R*, derived from this process and the initial conditions. One learns from such a model if *R* affects one's confidence in a hypothesis about the world. In the case of representational models, this may be because *Q* and *P* are sufficiently similar to a target to consider *R* relevant for that target, and hence information contained in *R* relevant for the confidence one has in hypotheses about the target. In the case of non-representational models, this may be because *Q* and *P* are at least considered possible, plausible or credible enough to consider *R* a relevant possibility. Considering *R* a relevant possibility then may affect one's confidence in certain hypotheses.

What part of a model is considered merely possible (rather than actual) and what kind of possibility is meant here will crucially influence whether the model result is considered a relevant possibility. It is therefore helpful to analyse different model types by the different possibility claims they contain. Here the extant literature on *how-possibly explanations* is very instructive. This literature controversially discusses what characterises how-possibly explanations, what distinguishes them from how-actually, potential, or how-possible explanations, and whether how-possibly explanations are explanations at all. In this paper, I eschew these controversies. Instead I use the conceptual distinctions offered by this debate to categorise different kinds of models, and to elicit the purposes and contexts in which the respective model types might offer learning opportunities.

The debate commences with Dray's (1957) claim that how-possibly explanations have a different aim and a different structure from how-actually explanations. How-possibly

explanations aim at giving an account how events that are considered impossible could have happened. How-actually explanations, in contrast, aim at accounting for how actual events have happened. Furthermore, Dray argues that how-possibly explanations rebut the impossibility of the explanandum by giving a necessary condition for its occurrence. He contrasts this with actual explanations offering sufficient conditions for their explanada. Reiner (1993) has criticised Dray's account, pointing out that how-possibly explanations do not really identify necessary conditions of the explanadum, but rather necessary parts of a sufficient condition for the explanadum.

This distinction is relevant for the present analysis. Actual explanation requires the identification of true (sufficient parts of) causes that brought about the explanadum. Representational models are one mode of identifying and representing these causes. How-possibly explanations, in contrast, identify elements of *possible* causes for an explanadum. Models can represent such possible causes – and hence contribute to how-possible explanations – without representing real-world targets. How-possibly explanations, in Dray and Reiner's sense, give non-representational models a purpose.

More recently, how-possibly explanations have been interpreted not in contrast to how-actual explanations, but rather as their precursors. According to this view, how-possibly explanations are similar to how-actually explanations, in that they satisfy most explanatory virtues, but they are inferior in that they lack adequate empirical support (Resnik 1991, 143). In particular, they are reasonably complete, showing how the explanadum was generated through a process from initial and background conditions. But process and background conditions are not well supported empirically, so that the account offers a mere possible, partial or potential explanation.

One may disagree whether Resnik's type should fall into the category "how-possibly explanation" (for a negative view, see Forber 2010). What is clear, though, is that non-representational model often serve the purpose that Resnik describes, and that this purpose is different from the one Dray and Reiner identify. First, models serving Resnik's type of how-possibly explanation will yield a result that represents a real-world target – otherwise, the similarity to how-actually explanations would not even arise. Models serving Dray-Reiner type how-possibly explanations, on the other hand, may yield results that do not represent real-world targets. Second, models for Resnik-type how-possibly explanations must be "reasonably complete" in order to be turned into how-actual explanations when empirical evidence for their similarity to some real-world target is forthcoming. No such requirement is imposed on models for Dray-Reiner type how-possibly explanations. They may serve their purpose of rebutting impossibilities with a rather sketchy structure, singling out only certain possible processes or background conditions.

Dray type how-possibly explanations focus on identifying some *conditions* that show the possibility of the explanadum. Another kind of how-possible explanation instead focus on indicating the sort of *process* through which the explanadum took place (Reiner 1993). Consecutive authors point out that this may consist in a mere proposal of a possible mechanism, or alternatively in providing a partial mechanism that in fact had the explanandum as outcome. In the latter case, the actual mechanism that produced the explanadum is identified, but in a way insufficient "to see more how the explanadum phenomenon was produced" (Persson 2011). Both purposes are served by non-representational models – the first by a model presenting a possible process, the second by a model presenting an actual process without sufficient causal detail, under possible background conditions.

Finally, Forber (2010) distinguishes between *global* and *local* how-possibly explanation. Global how-possibly explanations account for the possibility that an idealised object has a certain property, produced by a possible process from possible background conditions. Their purpose is to investigate the capabilities of general model processes (Forber 2010, 33). Local how-possibly explanations, in contrast, account for the possibility of a real target object having a certain property, produced by a possible process from actual background conditions. Their purpose is to guide speculation on how a particular model process can produce actual target properties. Forber's distinction thus points to a difference between non-representational models with an abstract result, and those with a concrete result.

Let me summarise. Non-representational models have a number of distinct purposes, which have been discussed in the philosophical literature under the heading of "how-possibly explanation". As the analysis of some of the key controversies in this literature showed, this notion contains a number of disparate scientific objectives – some of them explanatory, some offering other forms of epistemic gain, some merely heuristic. Crucially, these different purposes are served by different kinds of non-representational models. These models kinds can be distinguished by the modalities of the model result, the model process and the initial conditions. Keeping things simple and merely distinguishing between actual and possible (non-actual) processes and initial conditions, and concrete and abstract model results, we get six different kinds of non-representational models.[2] In the next section, I discuss each of these six non-representational models at the hand of an example, showing how in particular situations and for particular purposes, one can learn from each.

[2] Excluding both the representational model with actual initial conditions, actual process and concrete model result, as well as the representational model with actual initial conditions, actual process and abstract model result.

4. Six Cases of Learning from Non-Representational Models

My preceding abstract discussion leaves many ambiguous cases – a model may contain, say, some merely possible initial conditions, and still represent the workings of an actual process producing some abstract actual property (as e.g. Mäki 2009 argues). Whether such a case is to be counted as a representational or non-representational model will depend on the interpretation of the intentions of the modeller and the objectives of the models' users. Instead of debating this in the abstract, it will be perhaps more fruitful to discuss the issue of learning from non-representational models at the hand of concrete examples.

In the following, I give examples for each of these six kinds of non-representational models. For each case, I identify contexts and purposes in which these respective models offer an opportunity to learn about the world.

i. Possible initial conditions, possible process, abstract result

Axtell and Epstein's (1996) *sugarscape* is a set of models consisting of agents with individual rates of metabolism and fields of vision, a two-dimensional (51x51 cell) grid which contains different amounts of sugar on each cell, and rules governing the interaction of the agents with each other and the environment. In every step agents look around, find the closest cell filled with sugar, move and metabolize. If their sugar level is below their metabolism rate, they die. Harvested cells grow back one unit of sugar per time period. Using this basic set-up, Axtell and Epstein construct a model of migration, where agents' maximum vision is 10, and all agents are initially clustered together in one rectangular block in the southwest the grid. The authors do not claim that either the initial conditions of the model or the processes established by the model rules represent any actual target; they thus propose a non-representational model with merely possible background conditions and process.

Axtell and Epstein's model produces "waves of migration": a group of agents move outward in north-easterly direction from the initial cluster. Only when this group has progressed a considerable distance does the next group follow them. Although they mention wavelike movements in some mammal herds and economic "herding" as target for other models, they do not argue that the result of their model represents any such actual case. Instead, their result is a mere abstract pattern that *might* be instantiated in the real world.

And yet, one might learn from this model. Axtell and Epstein write that the model

produced "a phenomenon we did not expect" (Axtell and Epstein 1996, 42). Then they analyse the waves as produced by the interplay of food search and consumption by agents, and the slow regrowth of sugar; and they analyse the northeast direction of the migration (a direction in which single agents cannot move) as produced by "a complex interweaving of agents" (ibid.). The model thus justifies reducing one's confidence in the hypothesis that waves of migration cannot arise from mere food dynamics or that they cannot go in directions single agents cannot move. Because such patterns might be instantiated in the real world, such hypotheses are hypotheses about the real world. Anyone who had high confidence in these hypotheses (like apparently the authors themselves) learned from this model.

ii. Actual initial conditions, possible process, abstract result

Schelling's (1971) *checkerboard model* produces an abstract pattern of spatial segregation that he claims can be found in many cities, but which is not associated with any concrete settlement or even type of settlement. Schelling produces this abstract result with two types of tokens, initially distributed randomly over a checkerboard. Tokens move according to an iterated rule until no more movements occur. The rule is this. For a given token, if more than half of the tokens on (Moore-) neighbouring fields are of a different type, then this token will move to another vacant field with less than half of the neighbouring fields occupied with tokens of the other type. Schelling neither claims this process to represent an actual migration process, nor the checkerboard to represent an actual neighbourhood. But he claims that the process is started by an actual initial condition, namely the (non-racist) preference of individuals not to be in the minority. It is the one aspect of his model that he seeks to connect with the actual world, citing behavioural examples from restaurants, clubs and classrooms. Schelling's checkerboard model thus is a non-representational model with many possible and one actual initial condition, a possible process and an abstract result.

We learn from Schelling's model because it shows the *possible production* of an abstract pattern (a segregation of the two types of tokens on the checkerboard) from possible and one actual background condition and a possible process. In the context of spatial residential segregation, where the abstract segregation pattern might be realised, this possible production result is of particular importance: until then it was widely believed that racist preferences were a necessary cause of segregation. Schelling's model shows that segregation patterns might be produced by another cause, which is an actual condition in many real-world populations: namely the preference no to be in the minority. The model result thus justified changing one's confidence in hypotheses about racist preferences being a necessary cause of segregation. Anyone who had high confidence in such hypotheses learned from Schelling's checkerboard model.

iii. Possible initial conditions, actual process, abstract result

Güth's (1995) *indirect evolutionary approach* offers a model of preference evolution, which produces preferences for reciprocity. The model starts with a population of agents, who have different preferences over objects of choice (e.g. consumption bundles or behavioural strategies). Agents' rational choices then are determined according to their preferences, so that different preferences lead to different choices. Depending on their choice (and the environment in which the choice is made), an agent will have greater or lesser reproductive success than other agents with different preferences and hence different choices. Assuming that preferences are inherited, differential reproduction of agents then leads to differential replication of preferences in the population. Clearly, the background conditions of this model, in particular the distribution of preferences in the population, and the differential reproductive success of certain choices, are mere possibilities. The process by which the model result is produced, however, is an actual process, namely natural selection through differential reproduction. It has clear instantiations both in the domains of cultural and biological evolution. The result – preferences for reciprocation – are only described in abstract terms, and Güth makes no attempt to link it to concrete real-world targets. Nevertheless, one can learn from Güth's model. It shows that preferences with certain abstract properties[3] can be produced through selection in non-actual circumstances. That is, anyone who with high confidence believed that reciprocation, fairness or trust cannot be adaptive traits has good reason to change his belief when confronted with this model.

iv. Possible initial conditions, possible process, concrete result

Ainslie's (2001) *feedback model of self-control* produces a concrete result: the *moderate* impulsivity of human choices in the absence of precommitment devices, exemplified for example in the considerable number of addicts, most of whom eventually overcome their addiction. Ainslie produces this result with a possible description of delayed human value as an inverse proportion of delay, and a possible process of recursive self-prediction – prediction that is fed back to the on-going choice process. This description of value (also known as hyperbolic discounting) was first developed in order to account for impulsive choice, and hence is considered an actual initial condition by some. Yet the *moderate* impulsivity of human choice has led many to doubt that humans actually discount future value hyperbolically. It is exactly the aim of Ainslie's model to show that the hyperbolic description is compatible with moderate impulsiveness, by directly stoking it on the one hand, and by indirectly moderating it through a process of self-prediction that arises from this hyperbolic form itself. In the model, Ainslie thus intentionally casts the hyperbolic shape as a mere possibility. Furthermore, Ainslie readily admits that the process of recursive self-prediction is inaccessible to controlled experiment, and hence remains a

[3] In this case reciprocation, but in related papers Güth also produces preference for fairness and trust.

mere possibility.

Interestingly, in Ainslie's model, the proposed process of recursive self-prediction arises as a reaction to hyperbolic discounting, and it acts on future choices in the way often described as an effect of "the will" or "volition". Thus, one learns from Ainslie's model in two ways. First, the model justifies a change in confidence in the hypothesis that intertemporal behavioural data is incompatible with a hyperbolic shape of discounting. Second, the model justifies a change in confidence in the hypothesis that self-control can grow "from the bottom up" – from reactions to the hyperbolic shape of discounting. In Ainslie's words: "a small number of selected thought experiments yield a valid rejection of the null hypothesis – that contingent self-prediction is unnecessary for volition" (Ainslie 2009, 145). All those who had low confidence in such a claim learned from the failure of this model.

v. Actual initial conditions, possible process, concrete result

Axtell's et al. (2002) *Anasazi model* fails to produce a historically documented population dynamic of a settlement in the US southwest from soil and meteorological data, through any member of a set of possible migration decision processes of the modelled people. These possible decision processes involved rules whether to reproduce, to split up households, or to leave the settlement, given harvest levels. The model thus seeks to produce a concrete, actual phenomenon from actual initial conditions through a set of possible model processes. One can learn from this model by learning from its failure.

In particular, reference to the model justifies changing one's confidence in the hypothesis that the Anasazi's migration decisions based on subsistence considerations was sufficient to produce the exodus of the Anasazi around 1400 AD. Axtell's et al. model shows that with plausible processes, such a result cannot be produced from the actual conditions. Therefore, the model justifies increasing one's confidence in the belief that another capacity (cultural "pull factors" as the authors call it, in contrast to subsistence consideration "push factors") must be included in a model to produce the actual population dynamics from the initial conditions.

vi. Possible initial conditions, actual process, concrete result

Trivers (1971) *reciprocal behaviour model* produces a concrete actual result, the particular behavioural patterns exhibited by cleaner fish (*labroides dimidiatus*) and their hosts. To this end, it employs an actual process, frequency-dependent selection, which is found in many instances of biological and cultural evolution. Cleaner and host, so Trivers argues, are engaged in a indefinitely repeated Prisoners' Dilemma game, where the gains of cooperation (i.e. the cleaner cleans and the host does not eat the cleaner) are

sufficiently high to ensure differential reproductive success over unilateral defection. However, Trivers' model does not employ actual, but rather possible background conditions. In fact, the very purpose of Trivers' model is to identify initial conditions that would license a selection explanation of reciprocal behaviour between cleaner and host. These include:

> ". . . that hosts suffer from ectoparasites; that finding a new cleaner may be difficult or dangerous; that if one does not eat one's cleaner, the same cleaner can be found and used a second time; that cleaners live long enough to be used repeatedly by the same host; and if possible, that individual hosts do, in fact, reuse the same cleaner" (Trivers, 1971, 41).

That Trivers list these conditions in this way makes clear that his model is a non-representational model with merely possible initial conditions. Yet one learns from this model: it gives one good reasons to change one's confidence in hypotheses about what the necessary conditions are for reciprocal behaviour between cleaner and host to be an adaptive trait.

5. Conclusions

I have argued that one might justify non-representational models by showing that one learns from them about the world. I did not claim that one can learn from every non-representational model, and therefore that every non-representational model is justified. Instead, I described a possible way of appraising them, which is stronger than merely justifying them as heuristic tools.

To this end, I characterised learning as justifying a change in confidence in certain hypotheses about the world. I then discussed a number of hypotheses relating to possibility claims, and argued that changing one's confidence in any of them would affect the way scientists and policy makers seek to explain and control the actual world. These hypotheses, although relating to possibility claims, thus are about the world.

To analyse different kinds of possibility claims made with non-representational models, I employed conceptual distinctions from the discussion of how-possibly explanation. Six kinds of models emerged, distinguished by the modality of their background conditions, processes and results. Each of these kinds I illustrated with a concrete scientific model. In particular contexts and for specific purposes, I argued, one could learn from each of them. By demonstrating this, I showed that it is possible to justify each type of non-representational models, in particular contexts and for specific purposes. This concludes my argument.

References

Ainslie, G. (2001). *Breakdown of the Will*. Cambridge: Cambridge University Press.

Ainslie, G. (2009). Recursive self-prediction in self-control and its failure. In T. Grüne-Yanoff & S. O. Hansson (Eds) *Preference Change: Approaches from Philosophy, Economics, and Psychology*. Springer, 139-158.

Axtell, R. L., Epstein, J.M. (1996) *Growing Artificial Societies Social Science From the Bottom Up*. Cambridge, MA: MIT Press.

Axtell, R. L., Epstein, J. M., Dean, J. S., Gumerman, G. J., Swedlund, A. C., Harburger, J., Chakravarty, S., Hammond, R., Parker, J., & Parker, M. (2002). Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proceedings of the National Academy of Sciences, 99*(3), 7275–7279.

Dray, W. (1957). *Laws and explanations in history*. Oxford: Oxford University Press.

Forber, P. (2010). Confirmation and Explaining How Possible. *Studies in History and Philosophy of Biological and Biomedical Sciences 41*(1), 32-40.

Giere, R. N. (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Güth, W. (1995). An evolutionary approach to explaining cooperative behavior by reciprocal incentives, *International Journal of Game Theory*, Vol. 24: 323 - 344.

Hartmann, S.(1995). Models as a Tool for Theory Construction: Some Strategies of Preliminary Physics. In Herfel, W., Krajewski, W. Niiniluoto, I. and Wojcicki, R. (Eds.) *Theories and Models in Scientific Process*. Amsterdam: Rodopi, 49-67.

Hausman, D. M. (1992) *The inexact and separate science of economics*. Cambridge: Cambridge University Press.

Holyoak, K. and Thagard,P. (1995). *Mental Leaps. Analogy in Creative Thought*. Cambridge, Mass.: Bradford.

Mäki, U. (2009). MISSing the World. Models as Isolations and Credible Surrogate Systems. *Erkenntnis* 70 (1), 29-43.

Persson, J. (2011). Three conceptions of explaining how possibly - and one reductive account. In de Regt, Okasha, Hartmann (Eds.) *The European Philosophy of Science Association Proceedings*, vol 1. Dordrecht: Springer, pp. 275-286.

Richard Reiner *(*1993*)*. Necessary Conditions and Explaining How-*Possibly*. *Philosophical Quarterly* 44 (170):58-69.

Resnik, D. B. (1991). How-possibly explanations in biology. *Acta Biotheoretica* 39: 141-149.

Schelling T C (1971) Dynamic Models of Segregation. *Journal of Mathematical Sociology*, 1(2), pp. 143-186

Sugden, R. (2000). Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology*, 7(1), 1–31.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology,* 46, 35–57.

van Fraasen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.

Philosophy of Science

Inference to the Best Explanation and the Importance of Peculiarly Explanatory Virtues

--Manuscript Draft--

Manuscript Number:	11397
Full Title:	Inference to the Best Explanation and the Importance of Peculiarly Explanatory Virtues
Article Type:	PSA 2012 Contributed Paper
Keywords:	Inference to the Best Explanation; invariance; Woodward; Lipton
Corresponding Author:	David Harker, Ph.D. Jonesborough, TN UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	David Harker, Ph.D.
First Author Secondary Information:	
Order of Authors:	David Harker, Ph.D.
Order of Authors Secondary Information:	
Abstract:	Inference to the best explanation has at times appeared almost indistinguishable from a rule that recommends simply that we should infer the hypothesis which is most plausible given available evidence. In this paper I argue that avoiding this collapse requires the identification of peculiarly explanatory virtues and consider Woodward's concept of invariance as an example of such a virtue. An additional benefit of augmenting IBE with Woodward's model of causal explanation is also suggested.

Manuscript

Inference to the Best Explanation and the Importance of Peculiarly Explanatory Virtues

Abstract

Inference to the best explanation has at times appeared almost indistinguishable from a rule that recommends simply that we should infer the hypothesis which is most plausible given available evidence. In this paper I argue that avoiding this collapse requires the identification of peculiarly explanatory virtues and consider Woodward's concept of invariance as an example of such a virtue. An additional benefit of augmenting IBE with Woodward's model of causal explanation is also suggested.

1. Inference to the Best Explanation and the Threat of Vacuity

To illustrate the advantage of 'inference to the best explanation' (henceforth, IBE) over enumerative induction, Harman (1965, 90-1) invites us to consider inferences from samples to populations and the question of "when a person is and when he is not warranted in making the inference from "All observed A's are B's" to "All A's are B's"". Harman continues:

> The answer is that one is warranted in making this inference whenever the hypothesis that all A's are B's is (in the light of all the evidence) a better, simpler, more plausible (and so forth) hypothesis than is the hypothesis, say, that someone is biasing the observed sample in order to make us think that all A's are B's.

Clearly we can posit various reasons for why all the observed A's were also B's. It might be that "All A's are B's"; someone could have purposefully manipulated the sample to deceive us; perhaps our method for selecting subjects ensures, or makes it likely that, we will observe only those A's that are also B's, and so on. Furthermore, and equally patently, the actual reason for the observed regularity will be different in different cases. We observe only male drones, because all drones are male. Water that's pumped through an effective filter will contain no contaminants above a certain size; the absence of contaminants from the original water supply, however, often will not be the reason that the filtered water is pure. Harman supposes that such reasons can function as explanations. Let's concede that for now. Faced with competing explanations for an observed regularity Harman urges us to infer to the truth (or approximate truth) of whichever explanation is best.

Harman's proposal is thoroughly sensible – we should infer that hypothesis which is "better" and "more plausible".[1] However, without some guidance concerning how we identify the best, from competing explanations, and Harman has named a problem but not solved it. Insofar as IBE is regarded as a substantive theory of confirmation, its advocates can't rest content with an interpretation that advises only to infer that conclusion which is most plausible. Seemingly though Harman's phrase is sufficiently seductive, and has become sufficiently well-entrenched, that it is now hard to appreciate how vacuous the advice really is. Had Harman suggested we infer 'that hypothesis which seems most plausible in light of all available evidence', the attenuated condition of the suggestion would perhaps be more immediately apparent. If inferring to the best explanation is different, for Harman, it's hard to see how. On inspection, inference to the best explanation can appear quite insipid.

Lipton (2004), cognizant of the problem, offers a general means of responding. Unfortunately his development of that response opens him to critical objections, or so I'll argue in Section 2. The problems with Lipton's response trace to a failure to identify explanatory virtues, as distinct from virtues of the hypotheses that feature in the explanation. This diagnosis leaves room for a successful defense of IBE that utilizes Lipton's general strategy, but insists on peculiarly explanatory virtues, burdening advocates for IBE with the task of identifying such. Turning to the work of Woodward

[1] Harman does, in addition, suggest that better explanations are simpler, less ad hoc, and explain more. However, these concepts are insufficiently well-defined to provide helpful guidance in the face of competing explanations.

(for example, Woodward (2003)), I'll argue in Section 3 that distinctive explanatory virtues are apparent within the sciences and, furthermore, that it is not implausible to suggest that these reliably guide theory choice. Part of Woodward's project involves discriminating descriptions from explanations. An implication of this distinction is that Harman's example, above, might fall outside the scope of IBE, a possibility I discuss and welcome in Section 4. The purpose of the paper is not a complete defense of explanatory reasoning, but an attempt to motivate two important pieces of the groundwork: first, to urge that IBE requires the identification of explanatory virtues, and can't rely on the theoretical virtues of those hypotheses that are centrally involved in an explanation; second, to suggest that IBE has a limited scope, for purposes of understanding ampliative reasoning, which we might move some ways towards delineating by distinguishing descriptions from explanations.

2. Loveliness, Likeliness, Matching, Guiding

Concerned that IBE avoid appearing trite, Lipton responds in part by distinguishing two senses of 'best explanation'. The likeliest explanation, for Lipton, is that which is most likely to be correct. Informed that two theories each explain some phenomenon, we establish the likeliest explanation by evaluating which theory is best supported by available evidence. To infer to the likeliest explanation we needn't attend to anything about the explanations themselves; it is the well confirmedness of the respective theories that matters. The loveliest explanation, in contrast, can't be determined by attending to the merits of the underlying theory. Lipton suggests that the loveliest explanation "provides the most understanding". White (2005), endorsing Lipton's distinction,

suggests that explanations are often valued for "the degree of satisfaction" they deliver; explanations might disappoint because they are implausible, but also and alternatively because they can be "deeply unsatisfying". Having made this conceptual distinction, Lipton and White each suggest that IBE is a potentially important tool for investigating inductive reasoning, because explanatory loveliness might prove a reliable guide to explanatory likeliness. If this connection between loveliness and likeliness is real, we could justifiably appeal to the loveliness of an explanation for purposes of defending conclusions about which theory or hypothesis is most plausible, at least in some circumstances.

One concern with the proposal, as described, is that the concepts of understanding and satisfaction threaten to introduce a worryingly subjective dimension. What helps one person understand some phenomenon might differ from what helps another; explanations satisfy some folks, but not others. Judgments about differences in explanatory quality that ride on these kinds of consideration are unreliable markers of underlying plausibility. Lipton at least is careful to distance himself from overly psychological interpretations of the relevant concepts, but we can avoid such connotations altogether since the basic distinction suffices. Explanations can be evaluated in terms of the plausibility of the theory that motivates them, or in terms of features that are peculiar to explanations and independent of associated theories. In what follows I'll use the phrase 'explanatory virtue' to denote the latter. IBE avoids the charge of triviality by distinguishing explanatory virtues from the overall merits of a theory, and defining the rule as an inference based on the former; the plausibility of the rule, at least if it's understood

normatively, hinges on whether explanatory virtues reliably guide us towards a proper evaluation of available theories.

In furtherance of his claim that explanatory virtues need not be subjective, Lipton suggests simplicity, provision of mechanisms, scope, precision, among others, as appropriate measures of explanatory loveliness. None are unproblematic concepts, as Lipton concedes. Nevertheless, attaching loveliness here helps remove any lingering specter of subjectivity. Barnes (1995) protests, however, that these are not reliable guides to underlying plausibility. Suppose we have two competing explanations, but only one provides a mechanism. Whether we prefer the mechanistic explanation depends on the independent plausibility of the mechanism, suggests Barnes, rather than any intrinsic value in describing mechanisms. Lipton offers no obvious means of evaluating mechanistic hypotheses, but providing them can't be a reliable means of improving an explanation, or choosing between competing explanations, because even contrived and outrageous suggestions about the underlying mechanism describe a mechanism. Barnes raises similar complaints against the other putative explanatory virtues that Lipton describes.

Against the first edition of Lipton's book Barnes objections seem pertinent. Lipton (1991) asserts that "mechanism and precision are explanatory virtues" (118), "unification makes for lovely explanations" (119) and suggests that elegance and simplicity are also qualities of explanatory loveliness (68). He further argues that by attending to these qualities we are typically, reliably directed to the most plausible hypothesis. Lipton is unfortunately silent, however, on the issue of how we should balance the pursuit of these various

virtues, which might pull in opposing directions. If each virtue is evaluated in isolation, then Barnes objections are critical: discriminating purely on the basis of the presence or absence of a mechanism, for example, will often warrant an implausible inference. If, on the other hand, Lipton intends us to weigh all explanatory virtues and reach an appropriate balance between them, then his failure to describe how this should be conducted leaves the account disconcertingly obscure. Lipton's earlier defense is either reasonably transparent, but implausible, or quite opaque. However, Lipton's defense shifts between the two editions of his book. In the more recent he argues explicitly for a correspondence between theoretical and explanatory virtues, then argues independently, and on empirical grounds, that we in fact use the latter to evaluate the former. What is discussed as "matching" and "guiding" in the later edition are not distinguished in the earlier. Lipton hereby implies that the likeliest and loveliest explanations will each provide the best balance of various virtues, although again Lipton provides no guidance on how we are to recognize the best trade-off. Given Lipton's new strategy it becomes hard to accuse him of proposing an unreliable rule of inference, since it's a rule that by definition should guide us towards that conclusion which best instantiates all those theoretical virtues that are typically assumed important. The problems with Lipton's new strategy lie elsewhere.

One prominent theme in Lipton's book is that IBE describes our inferential practices better than alternative accounts. Lipton claims such advantages over Bayesianism, hypothetico-deductivism and Mill's methods of causal reasoning. Deficiencies with each, in terms of how well they describe our inferential practices, suggest either their

replacement with IBE or, in the case of Bayesianism, augmentation with explanatory considerations. These comparative claims have been challenged. Rappaport (1996) defends Mill's methods against Lipton's concerns. Bird (2007) argues that Lipton's objections are largely ineffective against hypothetico-deductivism. Douven (2005) argues that Lipton says too little about how and why Bayesians should build explanatory considerations into their framework. Furthermore, even if we concede that IBE better describes our inferential tendencies, we don't thereby achieve any normative justification for explanatory reasoning. What Lipton does say about the normativity of the rule is uninspiring.

According to Lipton's matching claim, explanatory reasoning is justified since explanatory considerations direct us towards that hypothesis which is most precise, has greatest scope, and so on, which Lipton suggests render that hypothesis most probable. However, Lipton offers little by way of analysis for these theoretical virtues. Consequently, because they're notoriously vague, and because it's hard to justify why they matter for purposes of confirmation, and because we don't know how to balance these often competing qualities against one another, Lipton leaves many hostages to fortune. The justification for explanatory reasoning is entirely derivative, and it is derivative on something that's worryingly vague. There is no answer as to why we should value a rule that directs us towards the simplest hypothesis, other things being equal. However, we might reasonably expect that if a theory of confirmation is going to place a premium on considerations of simplicity, then it should justify that decision. Leaving so

many concepts unanalyzed might leave us again wondering whether there's any real substance to IBE.

The failure to more carefully define these concepts becomes problematic again when we turn to Lipton's guiding principle. It is suggested both that, as an empirical matter, we tend to be impressed by explanatory considerations and, when confronted with competing explanations, it is the simpler, more precise, and so on, that is inferred. However, there is no obvious reason to suppose that the sense of simplicity that I employ when making a judgment about competing explanations will be the same sense that might prove a justified means of adjudicating between competing hypotheses.[2] A normative justification for Lipton's account requires either that we offer distinct analyses of explanatory and theoretical simplicity, then argue that explanatory simplicity is a reliable guide to theoretical simplicity, or we stipulate that simplicity has the same sense in each context. The former strategy is far from straightforward. The latter makes it much more difficult to argue that we in fact prefer simpler explanations, in the relevant sense and other things

[2] For example, in curve-fitting problems it has been argued that introducing additional adjustable parameters is appropriate only if will improve the predictive accuracy of the curve. If we define simplicity in terms of the number of adjustable parameters, then we justify a role for simplicity within certain well-defined contexts (see Forster and Sober (1994)). However, the balance between fit and number of parameters emerges from a non-obvious mathematical theorem. It seems unlikely that any 'intuitive' sense of simplicity that we might employ in evaluating explanations should guide us towards hypotheses that are more simple in this respect.

being equal. Maintaining both the guiding principle and a normatively justified interpretation of IBE becomes less plausible.

Hopes of preserving the normative dimension of IBE are further degraded when Lipton appeals to data from cognitive psychology. For example, Lipton describes the results of work conducted by Kahneman and Tversky, which demonstrated our propensity for committing the conjunction fallacy. (Asked to identify which event was most probable, given some scenario, many subjects committed the error of supposing a conjunction of two events can be more probable than one of the conjuncts.) Lipton offers this as evidence both that we are not good at Bayesian updating and that explanatory considerations play an important role in how we reason. An obvious concern is that Lipton's interpretation of the result provides an immediate example of explanatory reasoning that is unreliable. Lipton responds that in circumstances more complicated than those described by Kahneman and Tversky explanatory reasoning might be more reliable, but offers no evidence to support the conjecture.

In summary, Lipton argues that explanatory loveliness is both a reliable guide to explanatory likeliness, because considerations like simplicity and scope are features of more probable hypotheses and more virtuous explanations, and an important aspect of our inferential practices. However, the connections between these theoretical virtues and the plausibility of a given hypothesis are sufficiently vague that it is hard to admit them into a theory of confirmation as brute facts. The argument also requires us to concede that our natural proclivities, when evaluating explanations, will draw on similar considerations to those that will ultimately be deemed important for evaluating hypotheses, and that we

apply them in similar ways. Finally, in light of our demonstrated cognitive failures where we are perhaps unduly influenced by explanatory considerations, we must hope for evidence that such failures are heavily restricted to certain kinds of case. Absent such evidence and, although we might have reason to suppose we in fact employ explanatory reasoning, we'd lack any reason to suppose that we should. The normative dimension of IBE, as developed by Lipton, is both vague and tenuous. Admittedly Lipton at times seems content with defending a purely descriptive interpretation of IBE, in which we declare only that explanatory considerations in fact feature prominently in our reasoning. Typically IBE is understood as a normative thesis; a purely descriptive thesis certainly falls short of my ambitions for the rule.

Where did Lipton go wrong? I suggest it's in arguing that explanatory and theoretical virtues align. By adopting that position it becomes hard for explanatory considerations to illuminate, account for, or justify judgments about which of competing hypotheses is most plausible. The promise of IBE, as initially presented by Lipton, was with the idea that we could read off qualities of an explanation and thereby learn something important about the merits of the underlying theory. Given the matching claim, any normative justification for IBE becomes fully dependent on concepts that are not only problematic and vague, but also appear independent of explanatory considerations. Consequently, Lipton is forced to adopt an essentially descriptive interpretation of the rule. A model of IBE would be more useful and more interesting if we could identify peculiarly explanatory virtues, that cannot be identified with qualities of the underlying hypotheses, and that help us understand why certain inferences are sensible. Developed in this way

and IBE could live up to its reputation as a theory of how we should reason. Utilizing Woodward's model of causal explanation I'll now sketch a way of relating explanatory considerations to underlying plausibility that seems promising.

3. Invariance, Mechanisms and Consilience

Woodward's model is centrally concerned with change relating regularities, regularities that describe how changes in the value of one variable affect the value of another. Interventions on variables pick out causal and explanatory relations, for Woodward, if they are a reliable means of manipulating other variables within the regularity. Many regularities will satisfy this standard under some conditions but not others. For example, the ideal gas law properly captures our ability to increase the temperature of a gas by increasing the pressure, in certain circumstances. The law is thus a change-relating regularity that describes a causal relation, exploitable for purposes of explaining. The law doesn't hold universally, however. When temperatures become sufficiently low, or pressures sufficiently high, the law no longer accurately describes the relation between these variables. In such conditions we might appeal to the van der waals equation, which holds in circumstances where the ideal gas law breaks down. For Woodward, the latter is more invariant. Regularities are invariant if they continue to hold despite interventions on the variables that feature in that regularity. We explain an outcome by appealing to a system of regularities that is invariant under at least some interventions, and which can be combined with a range of possible initial and boundary conditions to describe how events would have differed had those conditions been otherwise. Only regularities that are invariant under some interventions are explanatory. Regularities that are more invariant

support a broader range of explanations, since they allow us to say more about how things would have been different if initial or background conditions were different.

Although Woodward isn't concerned with the relationship between invariance and confirmation, and even expresses some skepticism about inference to the best explanation (see note 5), I suspect there are important connections. My proposal is that it is reasonable to infer more invariant explanations, over less invariant explanations, because considerations of invariance tell us something important about the regularities that ground the explanations. My suggestion is that pursuing greater invariance will tend to produce the kinds of achievements that scientists consider epistemically significant, including our admiration for verified novel predictions, predictive success more generally, and high precision testing, our suspicion of ad hoc hypotheses, desire for both `deeper' explanations and explanations of `free parameters', as well as our pursuit of theories that have greater consilience. Despite their reputations, these concepts are poorly understood. The concept of invariance, insofar as it can illuminate these more familiar concepts, advances our understanding of confirmation.

Before offering some details, a few preliminaries are in order. First, invariance is distinct from predictive success, consilience, scope, and so on. The proposal thus shares with Lipton's defense a distinction between two types of explanatory achievement. We can evaluate an explanation in terms of its invariance, where more invariant explanations are better. Explanatory hypotheses and regularities can also be better insofar as they are less ad hoc, more precise, verified by novel predictions, and so on. If invoking the concept of invariance offers more plausible analyses for the confirmatory significance of such

considerations, then it has importance for our understanding of confirmation as well as explanation. What distinguishes my proposal from Lipton's more recent defense is that invariance is a peculiarly explanatory virtue, rather than a feature of the underlying theory or hypothesis. This creates room for a normative defense of explanatory reasoning. It is also important to distinguish a more modest from a more ambitious version of the thesis I'm proposing. The more modest rests content with providing a better account for extant confirmatory considerations. The more ambitious version assumes, or argues, that those concepts are in turn indicative of more general forms of scientific achievement. If pursuing invariance helps us achieve deeper explanations, for example, and deeper explanations indicate a more truthlike theory, then we connect a distinctively explanatory virtue to perhaps the ultimate scientific achievement. Admittedly concepts like consilience and ad hoc-ness are only poorly understood, thus difficult concepts to offer in defense of realist commitments. However, insofar as IBE might help provide more convincing analyses for various intuitions surrounding questions of confirmation, once augmented with Woodward's concept of invariance, it can simultaneously help justify its own normative credentials. It's beyond the scope of this paper to start properly exploring the connections between invariance and all the concepts I've alluded to. Hopefully a couple of examples will provide adequate motivation for the thesis.

First, let's return to Lipton's desire for mechanistic explanations and Barnes' concern that merely adding a mechanism can't itself reliably improve an explanation. The concept of invariance enables us to distinguish mechanisms that improve our explanations from those that don't. Drawing on Woodward's example, the amount of pressure applied to the

gas pedal explains the speed of my car, at least under some conditions. This change-relating regularity can be exploited for purposes of manipulating the speed of the car, and therefore for purposes of explaining the speed, even for those of us who are ignorant about how changing the pressure applied to the pedal brings about the change in speed. Providing a mechanism that relates these variables will not always produce a better explanation: fanciful mechanisms that have no grounding in experience describe mechanisms. Mechanisms which are more invariant than the crude regularity we begin with increase our ability to manipulate and control the speed of the car under a wider range of conditions. We improve our understanding of the counterfactual dependencies that describe the system. Providing a mechanism that relates distinct variables will improve an explanation only if it is more invariant than the regularity alone.

Providing mechanisms for causal regularities is an important scientific pursuit. Thoroughly speculative mechanisms, however, are not valued, requiring us to find means of distinguishing speculative from plausible mechanisms. The concept of invariance achieves that. Furthermore, it's at least plausible to suppose that this improved ability to manipulate a system reflects a better understanding for how a given system behaves.[3]

[3] Several authors have suggested that IBE has importance for purposes of fixing prior probabilities, likelihoods, or both, within Bayes' equation (for example, Lipton (2004), Okasha (2000), Weisberg (2009)). The rule is thus given a probabilistic interpretation. Elsewhere I've argued that advocates for this approach are vulnerable to a critical

As a second illustration, again inspired by Woodward (2003, 261-2), consider the puzzle of distinguishing consilience from conjunction. Conjoining two theories produces a new, more general theory. However, explaining events by appealing to a conjunction is no improvement over an explanation that appeals to the relevant conjunct. Conjoining Hooke's law with the ideal gas law doesn't improve our explanations for the temperature of a given gas, even though the conjunction is more general. Theories are, however, lauded for their consilience. Newton's theory of universal gravitation offered explanations for falling bodies, planetary motions and tidal effects via a unified system. Consilience involves more than just conjunction, but identifying the excess has proved problematic. Again the concept of invariance is edifying. Conjunctions provide no additional information about the effects of intervening on variables, beyond what's provided by one of the conjuncts in isolation. Frequently cited cases of consilience, in contrast, do provide additional information. Galileo offered explanations for bodies falling near the Earth's surface. Newton also offered explanations for bodies falling near the surface of Earth (or any other massive object), but his were invariant under changes to the mass and radius of the body on which the objects are dropped. Newton's explanations are invariant in ways that Galileo's are not. The concept of invariance accounts for the differing attitudes towards conjunction and consilience.

The concept of invariance promises valuable analyses of various confirmatory concepts. A convincing defense of this claim requires both a more careful explication of the two

---

dilemma and that IBE should instead be understood as a guide to better representations of target systems (see author).

concepts already presented, and their relation to invariance, and extended discussions of the other concepts I've alluded to. A satisfactory treatment lies beyond the scope of this paper, but hopefully I've done enough to at least induce some goodwill for the idea. Rather than develop this aspect of the project further, in the following section I'll explore an independent reason to regard Woodward's theory as a helpful crutch for IBE.

4. Descriptions, Explanations and IBE's Scope

For Woodward, explaining involves communicating relations of counterfactual dependence. Regularities that don't capture such relations can't be utilized for purposes of explaining, although they might provide useful and accurate descriptions of target populations. For example, "All swans are white" cannot explain why a particular swan is white, since it doesn't provide the kind of dependency to which Woodward attaches significance. The explanatory impotence of certain regularities has an important consequence for Harman's puzzle, described above. Concerned to identify those circumstances when it is appropriate to infer 'All A's are B's' given that 'All observed A's are B's', Harman suggests the inference is justified if the former provides the best explanation for the latter. If the regularity is not change relating however, then it doesn't explain at all, at least according to Woodward.

IBE is understood differently by different authors. One disagreement concerns the rule's scope. Harman (1965) and Psillos (2002) suggest the rule is more general than inductive reasoning; Lipton (2004) describes IBE instead as one important type of non-deductive reasoning. I favour Lipton's more modest attitude; some of the considerations that persuade me will be presented below. Adopting Lipton's position burdens one with

providing criteria for when IBE can, and cannot, be employed, and an intriguing platform for that project is precisely the distinction between descriptions and explanations that Woodward's model of explanations articulates. Sometimes our concerns are principally with describing a process, or kind; sometimes our concerns lie with explaining why certain events occurred, or why things are configured in a particular way. Restricting explanatory inferences to those circumstances when we are actually engaged in explaining seems sensible. It also helps insulate the rule against important objections.

Consider Hitchcock's (2007) objection, in which we imagine two coins, one fair and one biased (3:1) in favour of heads. A coin is selected at random and flipped four times, where each flip lands heads. We assume a prior probability of 1/2 that we selected a particular coin, conditionalize on the new evidence, and thereby determine the posterior probabilities. We know how probable it is that we selected either coin, but Hitchcock sensibly asks what reason IBE can offer for preferring one hypothesis over the other. Relative to the evidence, neither hypothesis is simpler, more unifying nor, more generally, more lovely. Thus while the Bayesian can give clear directives concerning which hypothesis is more probable, and by how much, advocates of IBE seemingly have little to offer. Hitchcock's concern is well-directed, but might serve to motivate the delineation described above. Whether the selected coin is fair, or not, is a question about whether we have properly described the propensity of the coin. Such descriptions will align more or less probably with the outcome of subsequent sequences of flips, which are thereby entirely relevant for purposes of evaluating the plausibility of the competing descriptions. By restricting IBE to the evaluation of change relating regularities, however,

the example falls outside the domain of IBE. Hitchcock is thus quite correct, I'd submit: IBE has nothing to offer in terms of illuminating such cases. The lesson is not that IBE is flawed, but that it has a restricted range of application.[4]

5. Conclusions

Inference to the best explanation faces various objections and would benefit from additional work along several dimensions. Most urgent, to my mind, is that the rule distinguish itself from a recommendation simply that we infer that conclusion which is most plausible given available evidence. A second significant challenge emerges from some very sensible criticisms: explanatory considerations are not always relevant to inductive reasoning, so the rule must have a more limited scope than some have suggested. The challenge is to identify those circumstances when IBE helpfully and properly models good inferential habits. In Woodward's account of causal explanation I've suggested that we may have the resources both to develop a potentially instructive and plausible version of IBE, and simultaneously start to better understand its boundaries.

[4] Woodward (2003, 5) also expresses doubts about IBE, arguing that the distinction between explanation and description is essential to a proper understanding of scientific methodology, but that descriptions are evidently not confirmed by appeals to explanatory qualities. Clearly, however, once we rescind hopes of developing IBE into a universal model of confirmation, Woodward's concern disappears.

References

Barnes, Eric. 1995. "Inference to the Loveliest Explanation." Synthese 103:251-77.

Bird, Alexander. 2007. "Inference to the Only Explanation." Philosophy and Phenomenological Research 74:424-32.

Douven, Igor. 2005. "Wouldn't it be lovely: Explanation and Scientific Realism." Metascience 14:331-61.

Forster, Malcolm and Elliott Sober. 1994. "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." British Journal for the Philosophy of Science 45:1-35

Harman, Gilbert. 1965. "The Inference to the Best Explanation." Philosophical Review 74:88-95.

Hitchcock, Christopher. 2007. "The Lovely and the Probable." Philosophy and Phenomenological Research 74:433-40.

Lipton, Peter. 1991. Inference to the Best Explanation. 1st edition London: Routledge.

Lipton, Peter. 2004. Inference to the Best Explanation. 2nd edition. London: Routledge.

Okasha, Samir. 2000. "van Fraassen's critique of inference to the best explanation." Studies in the History and Philosophy of Science 31:691-710.

Psillos, Stathis. 2002. "Simply the Best: A Case for Abduction." in Computational Logic: Logic Programming and Beyond: Essays in Honour of Robert A. Kowalski.

Rappaport, Steven. 1996. "Inference to the Best Explanation: Is It Really Different From Mill's Methods?" Philosophy of Science 63:65-80.

Weisberg, Jonathan. 2009. "Locating IBE in the Bayesian Framework." Synthese 167:125-43.

White, Roger. 2005. "Explanation as a Guide to Induction." Philosophers' Imprint 5:1-29.

Woodward, James. 2003. Making Things Happen: A Theory of Causal Explanation. Oxford University Press.

From Desire to Subjective Value:
On the Neural Mechanisms of Moral Motivation

Abstract

Increasingly, empirically minded moral philosophers are using data from cognitive science and neuroscience to resolve some longstanding philosophical questions about moral motivation, such as whether moral beliefs require the presence of a desire to motivate (Humeanism). These empirical approaches are implicitly committed to the existence of folk psychological (FP) mental states like beliefs and desires. However, data from the neuroscience of decision-making, particularly cellular-level work in neuroeconomics, is now converging with data from cognitive and social neuroscience to explain the processes through which agents are moved to act on the basis of decisions, including decisions about social and moral norms. I argue that these developments are beginning to cast doubt on the prospect of finding *nontrivial* physical 'realizers' for the FP states invoked in the Humeanism dispute by posing two distinctive challenges that tend to work against each other: belief-desire directionality and causal relevance.

> *A great deal of the recent work in cognitive science has, tacitly or explicitly, assumed very much the picture of mental organization that folk psychology proposes. There are other straws in the wind, however. There are findings and theories suggesting that something is seriously wrong with the simple belief-desire structure implicit in common sense wisdom.*
>
> *Stich (1983: 230)*

§1. From Metaethics to the Neuroscience of Decision

For those of us who worry about moral philosophy's empirical commitments in the age of science it is perhaps encouraging that there is now at least one problem with deep roots in analytic ethics that is receiving much empirical treatment from philosophers.[1] That is the

[1]The last decade or so has seen a growing number of philosophers express concern over the proliferation of dubious empirical claims and assumptions in ethics. Darwall et al. observe in their overview of the last century of work in ethics that, "too many moral philosophers… have been content to invent their psychology or anthropology from scratch" (1997: 34-5). Doris and Stich have recently echoed that concern, arguing that philosophy's empirical complacency has discouraged scientists from "undertaking philosophically informed research on ethical issues" (2005: 115).

problem of explaining moral motivation (hereafter "MM"), or the apparent tendency of an agent to be motivated by her moral judgments, that particular class of mental states generally understood as an attitude of assent to normative claims.

The problem of MM encompasses (at least) two distinct though related philosophical disputes. Motivational internalists (or just "internalists" for our purposes here) argue that moral beliefs motivate *necessarily* while externalists deny this. And proponents of the so-called Humean theory of motivation (Humeanism) argue that moral beliefs are *insufficient* for motivating agents, since motivation requires in addition to a belief the presence of a conative state such as a desire. Anti-Humeans reject the Humean theory on the grounds that moral beliefs are themselves *sufficient* for motivation. Many endorse internalism or one of a few related ideas such as that moral beliefs are simultaneously desire-like ("besires"[2]) or that moral beliefs co-occur with or otherwise trigger the relevant desires. A growing number of naturalistically minded philosophers are turning to data from psychology, psychiatry, cognitive science, and neuroscience to help resolve these longstanding philosophical disputes about MM.

For example, Roskies (2003, 2006) argues that patients suffering from damage to the ventromedial prefrontal cortex (VMPFC) serve as counterexamples to internalism.[3] Kennett (2002; Kennett & Fine 2009) argues that clinical research with autistic persons and psychopaths favors the Kantian account of motivation, a form of anti-Humeanism according to which moral judgments are necessarily motivating. Prinz (2006) also uses data on psychopathology, though to argue for a Humean sentimentalist account of moral concepts according to which an agent's believing an action morally wrong amounts to her having a sentiment of disapprobation toward

[2] The term is due to Altham (1986).

[3] More recently Roskies has argued, with Schroeder and Nichols, that instrumentalism—a variation on the Humean theory, which holds that an agent is motivated when she forms beliefs about how to satisfy her pre-existing desires—"fits well with the neuroscientific picture" of motivational processes (2010: 106). Instrumentalism is similarly situated in the FP tradition.

it.

These philosophers are, I think, right to recognize the limits of traditional philosophical methods like conceptual analysis, intuition, and armchair reflection for elucidating the mechanisms of MM. They tacitly endorse the rather plausible idea that scientific data can catalyze progress on what is increasingly agreed to be (at least in part) an empirical question.

But there is another assumption at work in each of these approaches that is, I think, considerably less plausible. It is the assumption that scientific research will ultimately preserve the framework of commonsense psychology in which disputes about MM are couched. At the heart of that framework is an explicit commitment to realism about folk psychological (FP) concepts like belief and desire. Humeans, anti-Humeans, internalists and externalists alike are in dispute about the role that these states play in brining about MM. Each of these views presupposes that the right (or best) account of the relationship between moral judgments and motivation will preserve beliefs and desires (or something near enough). Eliminativism, instrumentalism, and skepticism about FP states are neither forms of anti-Humeanism nor externalism.[4]

But, for reasons I will suggest presently, we have more and more reason to doubt that such intentional states do, will, or could feature into our best scientific accounts of the mechanisms of judgment and motivation – at least *in the particular capacity that the vindication of FP realism requires*. Existing lines of research in cellular and social neuroscience are already converging toward an account of the causal mechanisms of moral cognition and motivation

[4] Eliminativism is not a form of anti-Humean because the latter theory holds not just that beliefs are insufficient for motivation—a claim which might seem compatible with the nonexistence of FP states—but also that motivation requires the presence of a *desire* (or related FP state). Eliminativism is not a form of externalism because it seems there is not much sense in the eliminativist's taking a specific position on the effects of undergoing nonexistent states. Stich has made a similar point in response to Dennett's instrumentalism, arguing, for example, that only real entities and not useful fictions can have causes and effects (1983: 244).

which neither invokes commonsense FP states nor appears likely to lend itself to accurate redescription in FP terms.

§2. Neuroeconomics: The Emerging Neuroscience of Decision

The last decade has seen the development of an influential research program, "neuroeconomics," which weds behavioral economics with experimental neuroscience. Its key insight is to use well-vetted theories from behavioral economics to contextualize neural data generated by subjects engaged in tasks of judgment and decision. Somewhat simplistically, behavioral economists can look to neuroscience to reveal the physiological constraints on real agents that sometimes lead them to violate the axioms of normative economic models, while neuroscientists can look to economic theory to help develop algorithmic models of decision-making. I will suggest that—insofar as we accept the idea that moral judgments are *decisions* about what it is right, or best, to do under such-and-such circumstances—the data emerging from this new discipline are likely to cast doubt on the tenability of philosophical theories of MM.[5]

Two once-independent lines of research are now converging on a model of human moral and social decision. Neuroscientists engaged in neuroeconomics continue to produce data which supports a *two-stage mechanism* for decision-making in our neural architecture while social and cognitive neuroscientists continue to show that the same neural networks and regions implicated in that two-stage mechanism are implicated in a subject's making judgments or decisions about moral and social norms.

[5] I think that this assumption is plausible for a variety of reasons. For example, note the cost at which the FP realist rejects it: any uncontroversial instance of an agent's making a moral decision (rather than a moral judgment) must be treated as irrelevant to disputes about moral judgment and hence to MM. This seems much less palatable than accepting the idea that making moral judgments is a lot like making decisions in moral contexts. There are many such reasons to grant the assumption, though for the sake of brevity I leave them for another time.

The two stages in the two-stage mechanism for decision-making are *valuation* and *choice*. In valuation, subjects assign subjective values (SVs) to individual goods or actions in a range of options. At the behavioral level, SVs can be understood as economic values calculated by quantifying the subject's choices relative to the alternatives.[6] At the neural level, it turns out that these SVs can be defined as the mean firing rates in action potentials per second of specific populations of neurons. These neural SVs are learned, represented, stored, and ultimately used to guide motor systems. In the second stage, choice, this neural information concerning the most highly valued item or action is implemented into motor pathways to guide physical action.

The neural process is thus very much like the process postulated by behavioral economists whose traditional models of economic choice explain decision making "as if" it involves choosing a highly-valued option from among an array of options represented in common currency (the "common currency hypothesis"). Neurophysiological data now indicates that decision-making at the neural level does indeed seem to involve common currency. That currency is SV: the responses of particular populations of cells, quantifiable in real numbers whose units are action potentials per second, to each of the items or actions available.

The neural pathways and regions implicated in valuation are the ventromedial prefrontal cortex (VMPFC) and striatum, while those implicated in choice are the lateral prefrontal and parietal cortex (Kable & Glimcher 2009). Recordings from cells in the VMPFC have contributed to the localization of valuation. Researchers have identified three types of neurons which respond to the values of individual goods on offer *regardless* of whether they are chosen (offer value neurons), to the values of goods and actions *actually* chosen (chosen value neurons),

[6] For example, if a monkey chooses reward A (e.g., apple slice) when paired with one 1B (e.g., one raisin), 2B (e.g., two raisins), and 3B, it is indifferent at a ratio of 4B:1A, and it chooses B when 6B and 10B are offered, then the value of 1A is roughly equal to the value of 4B [i.e.: V(1A)=V(4.1B)] and hence has a subjective value of approximately 4.

and to the chosen action itself (choice neurons; Padoa-Schioppa & Assad 2006, 2008). Similarly, these three types of neurons have been found in the caudate and putamen of the striatum where research indicates they track the values of actions (rather than goods; Samejima et al. 2005). What is perhaps most remarkable about these results is that offer value signals in these neurons correspond precisely to the common currency postulated by most "as if" models in economic decision theory.

Research on the neural architecture for the second stage, choice, has so far implicated the lateral prefrontal and parietal cortex. Much of this research is based upon work with the visuo-saccadic control system in the primate brain. Neuroscientists interested in sensory-motor control have studied this system extensively. It appears to provide the mechanism by which information concerning the chosen option, and not the unchosen options, is implemented in motor systems downstream from the valuation circuitry. Hence the explicit link with motivation.

The details are complex, but briefly the idea is that neurons in the lateral intraparietal area (LIP), frontal eye fields (FEF), and superior colliculus (SC) form a network for visuo-saccadic decision-making. Studies with monkeys on saccadic decision-making tasks have repeatedly shown that the firing rates of neurons in LIP and FEF increase as evidence accumulates that a visual response will result in reward. Interestingly, once those firing rates cross a threshold, a saccade is initiated (Shadlen & Newsome 2001). Further research has since indicated that this firing rate threshold represents a *value threshold for movement selection*: the saccade is initiated when its value reaches the preset threshold (Roitman & Shadlen 2002).

Much is now known about the mechanism through which SVs—the currency for choice —are learned and represented in the primate brain. Dopaminergic (DA) neurons in the midbrain encode a reward-prediction error (RPE), i.e., the difference between the outcome of an action

actually experienced and the predicted outcome of the action (Schultz et al. 1997). Research indicates that the firing rates of these DA neurons are *linearly* related to RPE as calculated by behavioral-level economic models (Bayer & Glimcher 2005).

As these lines of research elucidate the mechanisms behind choice in the primate brain, social and cognitive neuroscientists are revealing that the same regions, most notably the striatum and VMPFC are consistently implicated in tasks in which subjects are asked to make moral and social judgments (e.g., Greene & Haidt 2002; Moll et al. 2002; Koenigs et al. 2007). While much work remains to be done, it is perhaps fair to say that direct links between moral judgment and decision-making and the neurophysiology of decision and motivation have now been established.

We must recognize that these results, so long as they continue to withstand scientific scrutiny, already have serious implications for philosophers interested in vindicating traditional philosophical theories of MM. These philosophers must either finally put paid to the task of locating traditional FP concepts like belief and desire in our going scientific explanations for motivational processes (call this the "location project"), or they must content themselves with formulating theories in commonsense terms (i.e., speaking from *within* the perspective of moral agency rather than *about* it, cf. Blackburn 1998) and jettison appeals to scientific data and claims about empirical respectability. But philosophers who appeal to the sciences to support philosophical theories of MM have already conceded this latter project. It is they who owe some plausible account of how FP states might plausibly "supervene" on the neurophysiology.

§3. Locating Beliefs and Desires: Two Challenges for FP Realism

The central feature of valuation is the theoretical construct of subjective value (SV) and the

mechanisms through which SVs are learned. In general terms, the trouble for the location projecct stems from a tension between the cognitive-level FP story about an agent's subjectively valuing an item or action and the neurophysiological mechanisms upon which that story must supervene. In more specific terms, the problem for the FP realist is that (1) SVs "exist" – they are genuine *neural* entities, and (2) their contribution to decision and motivation processes—i.e., their explanatorily relevant characteristics and functions—pertain uniquely to the biophysical level. Here are those two points summarized by the leading neuroeconomist, Paul Glimcher:

1. "There is nearly universal agreement among neurobiologists that a group of neural systems for valuation has been identified" (2009: 511-12); and

2. There is a large body of data which supports the hypothesis that "learning mechanisms distributed through the basal ganglia and frontal cortex contribute to the construction of what we refer to as subjective value. These areas are hypothesized to learn subjective values, at a biophysical level, through the well-studied process of synaptic plasticity" (519).

We have just briefly reviewed some of that data in the previous section. I want to conclude by outlining what I take to be two of the most pressing challenges that this data appears to pose for the FP realist's location project.

3.1. The Challenge of Causal Relevance

The first challenge for the location project is to get beliefs and desires, whether construed as functional roles (e.g. Canberra Plan), sentences in the head (e.g. Fodor), metaphorical "directions of fit" (e.g. Anscombe 1957 and her many followers), etc., to supervene on the immediate causes of choice selection, namely the specific cellular and molecular configurations that result from synaptic plasticity in dopamine pathways and which make possible measurements of SV, *without jeopardizing their causal relevance in the second stage—choice—*

which is directly related to motor implementation.

There are several options in any choice context. What the data from neuroeconomics shows if anything is that making a choice is a matter of selecting from among a set of actions or items—*each* of which has a SV quantifiable in neural terms—a highly valued item and implementing information concerning that option in motor pathways. For the FP realist this process of selection must involve a set of beliefs and desires about the options on offer. The trouble is that if beliefs and desires are instantiated in this account at all then they must be instantiated in such a way as to represent the features of *each* of the items or actions on offer *and* without jeopardizing the role of beliefs and desires in the kinds of explanations for MM that Humeans, internalists, and their opposition are offering.[7]

Prima facie, the most plausible way for the FP realist to locate FP states in the neural account is to insist that desires are identical to or somehow constituted by SVs (or, again, the cellular/molecular configurations that make their measurement possible). On this account of location, for any given context of choice with more than one option the FP realist will have to claim that choice involves selecting from among competing levels of desire. A monkey faced with the choice of grapes, bananas and raisins is essentially faced with the task of selecting from among competing desires for each of the fruits, and perhaps chooses on the basis of beliefs about

[7] The point appears to be something of a neuroscience analog for some recent objections to the possibility of formulating a so-called "belief-desire law" which some functionalists suppose capable of explaining the relationship between intentional states in the theory-theory. Such a law might claim, for example, that "people do what they believe will satisfy their desires." In a notable objection, Gauker (2005) argues that there are no such laws. First he criticizes the "simple formulation" of the belief-desire law according to which people do what they believe will satisfy their desires by pointing out that "there is never just one thing people desire; they always desire a lot of things. They cannot do everything they think will satisfy all of their desires, because they cannot do all of those things at once" (126). This data suggests that it is not just a platitude that people desire lots of things and cannot do them all, but that it is a fact about our neural architecture that *even if* we could locate some scientific analog for desire, say as part of SV, its ubiquity in the context of choice would render it explanatorily inert anyway.

the quantities available. Two grapes, the monkey *believes*, satisfy its *desires* better than one raisin. Dopamine, synaptic plasticity, learning and so on are merely the neurophysiological mechanisms upon which the cognitive events supervene.

Crucially, though, this approach to location will jeopardize the explanatory relevance of beliefs and desires in the traditional disputes about MM. For example, Humeans claim that moral beliefs are insufficient for motivation because they require the presence of a desire (or similar conative state) to motivate. Anti-Humeans deny this, generally because they are drawn to some kind of motivational internalism. On the account of location just given, the Humean theory (or better, the spirit of that theory) will be true only trivially and its opposition simply a nonstarter. It is true in a manner of speaking that desires are required for motivation, but the point is trivial because desires are present to varying degrees in *each* of the options, *including those that are ultimately bypassed*. Moral beliefs about the nature of the possible options are insufficient for motivation because all such beliefs in the context of choice are insufficient for motivation. It is a platitude that desire (so understood) must be present for motivation precisely because in any real choice it is always present in its making a contribution to SV.

Neither does this result serve as evidence for internalist forms of anti-Humeanism which postulate "besires," i.e. states which are simultaneously belief-like and desire-like, since in any given decision *each* of the unchosen options will be motivationally inert despite each being the object of our besires (as it were).

The rapid proliferation of FP states in any attempt at location—which results from our having a rather hazy commonsense conception of precisely what kinds of entities they are and consequently no principled or reliable method for picking out their realizers—prevents them from contributing anything of value to explanations of MM couched in causal language. For

when we gloss these complex neurophysiological processes in commonsense FP terms we end up abstracting too far away from the mechanisms most immediately relevant to the explanation. Given the mechanisms of valuation, the claim that an agent chooses a particular (moral) course of action because she desires to is really just vacuous.[8]

3.2. The Challenge of Belief-Desire Directionality

I have just suggested that on the most obvious account of location, desires are identical to or somehow constituted by SVs. Now consider the role of beliefs on this same account. The FP realist can perhaps say that the monkey *believes* that each of its options carries a specific value in terms of its desirability. But that seems to conflate belief and desire in the traditional philosophical sense of the terms. Beliefs, philosophers tell us, are about objective states of affairs or facts, not representations of facts about our subjective experiences of desire.

More importantly, though, when we locate FP states in this way the interesting question is no longer whether desire must be present *in addition* to belief in order for choice and motivation to occur—which is the question contested by Humeans and anti-Humeans—but rather just the opposite: how beliefs *about* the desirability of an action or item contribute to selection.

To see this, consider first that it is a consequence of the MM framework and Humeanism in particular that we must find some role for belief as well as desire in the neural explanation. Finding neural correlates for desire at the cost of preserving any role for belief in a neural explanation for MM is hardly a victory for FP realism. But while it seems clear enough that for

[8] That is, at least, barring the development of an account of desire fine-grained enough to permit us to pick out only certain constituent components of SV rather than SVs themselves. But such fine-grained accounts of desire are not likely to be forthcoming for good reason. The more fine-grained the account becomes, the fewer commonsense cases of desiring it is likely to cover. FP realists are sensitive to this problem. Jackson and Pettit (1990) explain that the difficulty is to provide an explication of FP states that is specific enough to capture our commonsense attributions yet vague enough to render refutation by neuroscience unlikely. The challenges presented in this section are intended to illustrate why achieving this golden mean is likely impossible.

the FP realist desires must somehow be closely connected to SV, it is far less clear what role remains for belief, except perhaps to say that agents have beliefs about their subjective desires (SVs). This, though, yields the peculiar result that an agent navigates the world using representational desires and that her course of action is ultimately determined by the presence of a scale-tipping belief about which is the optimal desire to satisfy. That is, this particular account of location might find some room for both belief and desire only by turning the dispute about Humeanism in the wrong direction.

§4. Some Preliminary Conclusions

SVs, the neural common currency for choice selection, are unlikely to deliver the supervenience base that the traditional philosophical account of desire requires. It is difficult to see how the FP realist could provide an account according to which SVs constitute or realize desires without also realizing—entirely or in part—beliefs. SVs are neural representations of facts about the physical constitution of the *world*. Yet they are also neural representations of the facts about how that world, the physical environment, impacts the physical states of agent S's nervous system *in particular*. These complex neural representations are for this reason unlikely to admit of accurate description in the crude vocabulary of FP.

Are there alternative ways that the FP realist might locate commonsense states in this neural model of decision-making that could preserve causal relevance and directionality and ultimately vindicate commonsense theories of MM? It is, most will argue, far too early to rule out the possibility entirely. Even so, the emerging model of decision and motivation sketched here appears poised to deliver explanations of MM which differ from commonsense FP theories not merely in degree but in kind. The time is ripe to begin rethinking the commonsense

psychological framework upon which contemporarily analytic ethics is built.

References

Altham, J.E.J. (1986). The Legacy of Emotivism. In: Macdonald, G. & Wright, C. (eds.), *Fact, Science and Morality: Essays on A.J. Ayer's Language, Truth and Logic*. Basil Blackwell: 275-288.

Anscombe, G.E.M. (1957). *Intention*. Basil Blackwell.

Bayer, H.M. and Glimcher, P. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* 47: 129-41.

Blackburn, S. (1998). *Ruling Passions*. Oxford: Clarendon.

Casebeer, W.D. (2003). *Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition*. Cambridge: MIT Press.

Darwall, S., Gibbard, A & Railton, P. (1992). Toward *Fin de siècle* Ethics: Some Trends. *The Philosophical Review* 101: 115-189.

Dennett, D. (1987). *The Intentional Stance.* Cambridge: MIT Press.

Flanagan, O. (1991). *Varieties of Moral Personality*. Cambridge: Harvard University Press.

Gauker, C. (2005). The Belief-Desire Law. *Facta Philosophica* 7: 121-44.

Greene, J. and Haidt, J. (2002). How (and Where) Does Moral Judgment Work? *Trends in Cognitive Sciences* 6(12): 517-523.

Jackson, F., Pettit, P., & Smith, M. (2004). *Mind, Morality, and Explanation*. Oxford: Oxford University Press.

Kable, J. and Glimcher, P. (2009). The Neurobiology of Decision: Consensus and Controversy. *Neuron* 63: 733-45.

Kennett, J. (2002). Autism, Empathy, and Moral Agency. *The Philosophical Quarterly*, 52 (208): 340-357.

Kennett, J. and Fine, C. (2009). Will the Real Moral Judgment Please Stand Up? The Implications of Social Intuitionist Models of Cognition for Meta-ethics and Moral Psychology. *Ethical Theory and Moral Practice* 12: 77-96.

Koenigs, M., Young, L., and Adolphs, R. (2007). Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments. *Nature* 446: 908-11.

Moll, J., de Oliveira-Souza, R., Bramati, I.E., and Grafman, J. (2002). Functional Networks in Emotional and Nonmoral Social Judgments. *NeuroImage* 16: 696-703.

Padoa-Schioppa, C. and Assad, J.A. (2006). Neurons in the Orbitofrontal Cortex Encode Economic Value. *Nature* 441: 223-26.

Padoa-Schioppa, C. and Assad (2008). The Representation of Economic Value in the Orbitofrontal Cortex is Invariant for Changes of Menu. *Nature Neurosciene* 11: 95-102.

Prinz, J. (2006). The Emotional Basis of Moral Judgments. *Philosophical Explorations* 9 (1): 29-43.

Roitman, J.D. and Shadlen, M.N. (2002). Response of Neurons in the Lateral Intraparietal Area During a Combined Visual Discrimination Reaction Time Task. *Journal of Neuroscience* 22: 9475-9489.

Roskies, A. (2003). Are Ethical Judgments Intrinsically Motivational? Lessons from 'Acquired Sociopathy'. *Philosophical Psychology* 16(1): 51-66.

Roskies, A. (2006). A Case Study in Neuroethics: the Nature of Moral Judgment. In: J. Illes (ed.), *Neuroethics: Defining the Issues in Theory, Practice, and Policy*. Oxford: Oxford University Press.

Samejima, K., Ueda, Y., Doya, K. and Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science* 310: 1337-40.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.

Schroeder, T., Roskies, A., and Nichols, S. (2010). "Moral Motivation". In: J. Doris (Ed.) *The Moral Psychology Handbook*. Oxford University Press.

Schultz, W., Dayan, P. and Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. *Science* 275: 1593-99.

Shadlen, M.N. and Newsome, W.T. (2001). Neural Basis of a Perceptual Decision in the Parietal Cortext (area LIP) of the Rhesus Monkey. *Journal of Neurophysiology* 86: 1916-36.

Shadlen, M.N., Britten, K.H., Newsome, W.T. and Movshon, J.A. (1996). A Computational Analysis of the Relationship Between Neuronal and Behavioral Responses to Visual Motion. *Journal of Neuroscience* 16: 1486-1510.

Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge: MIT Press.

The Classical Continuum without Points

Geoffrey Hellman and Stewart Shapiro*

Abstract

We develop a point-free construction of the classical one-dimensional continuum, with an interval structure based on mereology and either a weak set theory or logic of plural quantification. In some respects this realizes ideas going back to Aristotle, although, unlike Aristotle, we make free use of classical "actual infinity". Also, in contrast to intuitionistic, Bishop, and smooth infinitesimal analysis, we follow classical analysis in allowing partitioning of our "gunky line" into mutually exclusive and exhaustive disjoint parts, thereby demonstrating the independence of "indecomposability" from a non-punctiform conception. It is surprising that such simple axioms as ours already imply the Archimedean property and that they determine an isomorphism with the Dedekind-Cantor structure of $\mathbb{R}$ as a complete, separable, ordered field. We also present some simple topological models of our system, establishing consistency relative to classical analysis. Finally, after describing how to nominalize our theory, we close with comparisons with earlier efforts related to our own.

.

1 Introduction

Since Aristotle[1], many mathematicians and philosophers have expressed the view that a genuine continuum cannot be composed of points. Related to this is the idea, also Aristotle's, that a true continuum is "seamless" or "indecomposable": it shouldn't be possible to break it apart cleanly, to "separate out" a proper part from the rest.

*Department of Philosophy, University of Minnesota (Hellman); Department of Philosophy, The Ohio State University (Shaprio). E-mail addresses: hellm001@umn.edu; shapiro.4@osu.edu.

Of course, the mainstream Cantor-Dedekind theory, along with the set-theoretic tradition, respects neither of these properties. But alternatives, such as the intuitionist and Bishop conceptions and smooth infinitesimal analysis ("SIA"), don't respect both of these either: they clearly have points. On the intuitionistic or Bishop constructions, the continuum is *entirely* made up of points, although these have to be determined by *constructive* Cauchy sequences of rationals, not arbitrary such sequences. SIA has points galore but it entertains (without asserting the existence of) nilsquare and nilpotent infinitesimals forming a "micro-neighborhood" $\triangle$ (of 0, translatable anywhere along the smooth line), behaving as a mini-line ("linelet") that can be translated and rotated but not bent, i.e. an axiom sitipulates that any (smooth) function defined on $\triangle$ obeys the equation of a straight line. (Its slope at any locus on the smooth line gives the derivative of the function there.)[1] Both these alternative approaches, however, do "respect" indecomposability, in that they forswear recognition of any function that would take, say, a constant value on one segment of the line and another constant value on the rest. And this, in turn, is achieved by giving up the law of excluded middle, i.e. by restricting the background logic to be intuitionistic. (The rationales offered by intuitionism and SIA, respectively, differ greatly, but the effects are quite similar in a number of respects, indecompsability counting as one of them.)

It would be wonderful to present Aristotle with these developments. We will leave it to scholars (and imaginative script-writers) to conjecture what he might have said. But the modern-day classicist, and perhaps many an impartial observer, might say that indecomposability is being "achieved" only in a negative sense, that is, by depriving oneself of the logical means of distinguishing—in the sense stipulated, viz. via total 2-valued functions—one part of the line (or respective lines they recognize) from the rest.[2] Our approach, developed in this paper, will decidedly *not*

[1] See Bell [2] for details on the development of SIA. Bell explicitly motivates SIA by expressions of dissatisfaction with point-based analyses of continua, and insists that the nilpotent infinitesimals of SIA are not to be conceived as further points. Just why not is not entirely clear, however. Perhaps it is because they (if they exist—something that cannot be proved or refuted in SIA) would be *too* indefinite as to location or extent to be thought of as points. Perhaps also they are collectively to be thought of as a kind of "glue" that holds the more definite points of the smooth line together.

[2] We are not saying that the restriction to intuitionistic logic is not well-motivated from the perspectives of intuitionism and SIA, but merely that indecomposability follows from the restriction in that it prevents recognizing any discontinuous functions. Indecomposability does not emerge from an analysis of the continuum and its constituents. It should be noted, however, that indecomposability takes different forms in intuitionism and SIA: the indecomposable subsets of the smooth line correspond

respect indecomposability; however, unlike all three approaches already mentioned, it *will* be truly non-punctiform. Points or numbers will be constructed (in several ways); however they will be clearly seen as "our additional superstructure" to a thoroughly non-punctiform line.[3] Moreover, in one version of this approach, points will only be recognized as "possible" additional structure, which seems quite in accordance with Aristotle's conception. Our approach may thus reasonably be called "semi-Aristotelian".

2 Atomless Mereological Continuum

The system developed in this paper is designed to characterize a one-dimensional continuum consisting of "regions" as parts, including "intervals"—although, as will be explained, the notions of 'open', closed', and 'half-open' are not available in our system. This continuum does not have any points as parts, although we will be able to define "point" in terms of intervals. Once we have proved that our continuum is Archimedean, we will demonstrate that it is isomorphic to the classical Dedekind-Cantor continuum, as a complete, separable, linearly ordered field.

Our formalism consists of classical first-order logic with equality supplemented with a standard axiom system for second-order logic (or logic of plural quantification, with an unrestricted comprehension axiom for plurals[4]), and with an adaptation of the standard (Tarski) axioms of mereology together with (something implying) the "Atomless" axiom.

Axioms of Mereology:

1a. Axioms on $x \leq y$ ("x is part of y"): reflexive, anti-symmetric, transitive.

Certain of our axioms and theorems are conveniently stated in terms of a binary relation called "overlaps": "*x overlaps y*": $x \circ y \Leftrightarrow^{df} \exists z(z \leq x \wedge z \leq y)$.

1b. Axiom on $\leq$ **and** $\circ$: $x \leq y \leftrightarrow \forall z[z \circ x \rightarrow z \circ y]$.

to a proper sub-class of the subsets of the intuitionist continuum indecomposable there. Cf, Bell [3].

[3] We postpone a comparison with more recent constructions along similar lines until the final section, below.

[4] This looks very much like second-order logical comprehension for monadic predicates, except that it is conditional upon there being something satisfying the predicate. It may be written:

$$\exists v(\Psi(v)) \rightarrow \exists xx \forall y[y \; \eta \; xx \leftrightarrow \Psi(y)],$$

where '$y \; \eta \; xx$' is read "y is one of (or is among) the xx", and where Ψ is any formula of the language lacking free 'xx'.

Theorem 1: *Axioms 1a and 1b imply the Extensionality Principle:* $x = y \leftrightarrow \forall z[z \circ x \leftrightarrow z \circ y]$.

Proof: From left to right is trivial. (Take y as x, then substitute y for the second x.) From right to left: Assume the right of Extensionality and rewrite it as the conjunction of two conditionals: $\forall z[z \circ x \rightarrow x \circ y] \wedge \forall z[z \circ y \rightarrow z \circ x]$. By Axiom 1b, the first of these yields $x \leq y$, and the second yields $y \leq x$. By anti-symmetry of $\leq$, the conjunction of these is equivalent to $x = y$.■

2. Fusion or whole comprehension: $\exists u\Phi(u) \rightarrow [\exists x\forall y\{y \circ x \leftrightarrow \exists z(\Phi(z) \wedge z \circ y)\}]$, where Φ is a predicate of the second-order language (or language of plurals) lacking free x.[5]

At this point, we could add an Atomless axiom: $\forall x\exists y(y < x)$, where $y < x \leftrightarrow^{df} y \leq x \;\&\; y \neq x$ (read "*y is a proper part of x*"). But this will follow from a stronger condition imposed below on the interval structure of our "pointless" or "gunky" line (axiom 5.).

We write $x + y$ for the mereological sum or fusion of x and y, such that $\forall z[z \circ x + y \leftrightarrow (z \circ x \vee z \circ y)]$, and we use $\sum_{n=0}^{\infty} x_n$ to designate fusions of infinitely many things. Also, if $\exists z(z \leq x \wedge z \leq y)$, then we write $x \wedge y$ for the *meet* of x and y, which satisfies $\forall z[z \circ x \wedge y \leftrightarrow z \circ x \wedge z \circ y]$. (If x and y have no common part, $x \wedge y$ is undefined.) And we write $x|y$ for $\neg\exists z[z \leq x \wedge z \leq y]$, pronounced "$x$ is discrete from y (and vice versa)". Furthermore, if $\exists z(z \circ x \wedge \neg(z \circ y))$, then $x - y$ is that part of x which does not overlap y, viz. $\forall z[z \circ x - y \leftrightarrow (z \circ x \wedge \neg(z \circ y))]$.(If there is no such z, then $x - y$ is undefined.) By axiom 2, fusions always exist, and meets and differences also exist wherever defined.

The pointless line we wish to characterize we'll label G, for "gunky".[6] Below, we'll prove that (quite remarkably) our very elementary axioms suffice to characterize G precisely as a certain minimal closure; and then

[5]The formulation in language of plurals takes this form:

$\forall uu\{\exists w(w \;\eta\; uu) \rightarrow \exists x\forall y[y \circ x \leftrightarrow$
$\exists z(z \;\eta\; uu \;\&\; z \circ y)]\}$,

where 'uu' is a plural variable, '$w \;\eta\; uu$' is read 'w is one of the uu'. (If plural variables are assumed to have instances, then the antecedent and the main conditional can be omitted.)

[6]The technical term 'gunk' for the "stuff" of atomless mereology stems from Lewis.[10, 1991]

Note that, by taking Φ in Axiom 2 as '$x = x$', a universal individual exists. Since, in what follows, we will find useful it to introduce a denumerable infinity of atoms to serve the role of natural numbers (i.e. the atoms with the usual operations defined on them collectively form an N-structure, i.e. satisfy the Dedekind axioms), we do not identify the universal object with G.

we'll prove that, with its interval structure, G is isomorphic to the classical real-numbers structure, $\mathbb{R}$.. The point, of course, of having the Atomless condition is to insure that, literally, G contains no points at all. Thus, except where we explicitly refer to atoms of an N-structure, the range of our first-order and plural variables can be thought of as all the parts of G, which we also call "regions".

It is convenient to introduce a geometric primitive, $L(x, y)$, to mean "x is (entirely) to the left of y". The axioms for L specify that it is irreflexive, asymmetric, and transitive. And we define '$R(x, y)$', "x is (entirely) to the right of y", as $L(y, x)$.

Now we can introduce an important geometric relation, *betweenness:* $Betw(x, y, z)$ for "y is (entirely) between x and z":

$$Betw(x, y, z) \Leftrightarrow^{df} [L(x, y) \ \wedge \ R(z, y)] \ \vee \ [R(x, y) \ \wedge \ L(z, y)]$$

It follows that $Betw(x, y, z) \leftrightarrow Betw(z, y, x)$.

$L(x, y)$ obeys the following axioms:

3a. $L(x, y) \vee R(x, y) \rightarrow x|y$**.** (Of course, $x|y$ implies $x \neq y$.)

3b. $L(x, y) \leftrightarrow \forall z, u[\ z \leq x \wedge u \leq y \rightarrow L(z, u)]$.

The following can now be inferred:

$Betw(x, y, z) \rightarrow x|y \ \& \ y|z \ \& \ x|z$, and
$Betw(x, y, z) \wedge Betw(u, x, z) \rightarrow Betw(u, y, z)$,

where the transitivity of L is used for the latter.

Now we can define an essential notion, that of a "*connected* part of G". Intuitively, such a part has no gaps. The definition is straightforward:

$$Conn(x) \Leftrightarrow^{df} \forall y, z, u[z, u \leq x \ \wedge \ Betw(z, y, u) \rightarrow y \leq x]. \quad (\text{Df } Conn)$$

("Anything lying between any two parts of x is also a part of x.")

Furthermore, we can define what it means for a connected part of G to be *bounded:* Let $Conn(p)$: then

$$Bounded(p) \Leftrightarrow^{df} \exists x, y[Conn(x) \ \wedge Conn(y) \ \& \ Betw(x, p, y). \quad (\text{Df } Bounded)$$

(A connected region wholly between two others is bounded.) Once we establish that G is bi-infinite, i.e. infinite in both directions, it will follow that, for connected regions, boundedness is a necessary condition for "*finite in extent*", as commonly understood. And once we have established that G is Archimedean, it will follow that boundedness is also sufficient for "finite in extent".

We call bounded connected regions "*intervals*" and write '$Int(j)$', etc., when needed. However, note that, lacking points, we cannot describe intervals as either "open" or "closed", or "half-open".

Using L, we can impose a condition of dichotomy for discrete intervals:

4. Dichotomy axiom: $\forall i, j[i, j$ are two discrete intervals $\rightarrow (L(i, j) \vee L(j, i))]$.

Now we can prove a linearity condition among intervals:

Theorem 2 *(Linearity)*: *Let x, y, z be any three pairwise discrete intervals; then exactly one of x, y, z is between the other two.*

Proof. Applying Dichotomy to the hypothesis, assume that (say) $L(x, y)$. If also $L(y, z)$, then $R(z, y)$, so that $Betw(x, y, z)$, and this is unique. If instead $L(z, y)$, then either $L(z, x)$, in which case we have $Betw(z, x, y)$, uniquely; or $L(x, z)$, in which case we have $Betw(x, z, y)$, also uniquely. The argument from assuming at first that $R(x, y)$ is similar. ■

To guarantee that arbitrarily small intervals exist everywhere along G, we adopt the following axiom:

5. $\forall x \exists j[Int(j)\ \&\ j < x]$.[7]

An important relation of two intervals is "adjacency", which is defined as follows:

$$Adj(j, k) \Leftrightarrow^{df} j|k \wedge \nexists m[Betw(j, m, k)]. \qquad \text{(Df } Adjacent\text{)}$$

Now suppose that $j = \Sigma_{i=1}^{\infty} j_i$, where $Int(j_i)$ and $R(j_{i+1}, j_i)$ and $Adj(j_{i+1}, j_i)$. Then we write $R(k, j)$ just in case $\forall i[R(k, j_i)$; analogously for $L(k, j)$. This will be useful in the proof of Theorem 3, below.

The following equivalence relations on intervals will also prove useful: "*j and k are left-end equivalent*" just in case $\exists p[p \leq j\ \&\ p \leq k \wedge \nexists q(\{q \leq j \vee q \leq k\}\ \&\ L(q, p)]$. "*Right-end equivalent*" is defined analogously.

One further geometric primitive is very useful both in insuring that G is infinite in extent and in recovering, in effect, the rational numbers as a countable, dense subset of the (arithmetic) continuum, viz. *congruence*, as a binary relation among intervals. Intuitively, $Cong(i, j)$ is intended to mean "the lengths of intervals i and j are equal". Thus, we adopt the usual first-order axioms specifying that *Cong* is an equivalence relation. We will sometimes write this as $|i| = |j|$, but with the understanding that we have not yet given any meaning to '$|i|$' standing

[7] This of course implies the "Atomless axiom", introduced above.

alone, but only in certain whole contexts. Similarly, for intervals i, j, we can define, contextually, $|i| < |j|$ as meaning: $\exists j'[j'$ an interval $\wedge\ j' < j \wedge \mathrm{Cong}(i, j')]$; and we may write $|i| > |j|$ as equivalent to $|j| < |i|$. Our next axiom will guarantee that these comparisons make general sense.

We come now to a key axiom, crucial to our characterization of G:

6. Translation axiom: Given any two intervals, i and j, each is congruent both to a unique left-end-equivalent and to a unique right-end-equivalent of the other.

In effect, this guarantees that a given length can be "transported" (more accurately, instantiated) anywhere along G, and that these instances are unique as congruent and either left- or right-end equivalent to the given length. In particular, we can prove

Lemma 1 *Given any two intervals i and j such that $\neg Cong(i, j)$, either there exists an interval $i' < j$ with $Cong(i, i')$;or there exists i'* with *$j < i'$ with $Cong(i, i')$.*

Proof. By $\neg Cong(i, j)$, $i \neq j$. Assume that $\neg(i' < j)$ for any i' such that $Cong(i, i')$. By the Translation axiom, there exists i' such that $Cong(i, i')$ and i' is left-end-equivalent to j. We want to show that $j \leq i'$, as that will establish that $j < i'$, as desired. Assume the contrary, i.e. that $j \not< i'$. Now, if $i' - j$ doesn't exist, then, by definition, $i' \leq j$. But, since $Cong(i, i')$, we have $\neg Cong(i', j)$, whence $i' \neq j$, and then we would have $i' < j$, contrary to hypothesis. So assume some $n \leq i' - j$. By the hypothesis for *reductio*, there is also $k \leq j$ & $k \not< i'$, and indeed $\neg(k \circ i')$. Without loss of generality, we may assume that k is an interval. (See axiom 5.) Since i' is left-end-equivalent to j, it follows that k is *not* left-end-equivalent to j. But there is $m \leq j - k$ which is left-end equivalent to both j and i' so satisfies $m \circ i'$. Let m' be a common part of m and i'. Clearly $L(m', k)$. But $\neg(L(n, k))$, since if it were, it would overlap j, contrary to assumption. (n can't be left of j, since it's part of i' and $i'$and j are left-end equivalent.) Therefore, by the Dichotomy axiom on L, we have $L(k, n)$, whence $Betw(m', k, n)$, with both $m', n \leq i'$ but $k \not\leq i'$, contradicting that i' is an interval. ■

Theorem 3 *(Trichotomy) For any two intervals, i, j, either $|i| = |j|$ or $|i| < |j|$ or $|i| > |j|$.*

Proof. Immediate from Lemma 1 and the definitions of the disjuncts. ■

One further axiom on congruence is useful and intuitively intended, viz. that congruence respects nominalistic summation of adjacent intervals:

7. Additivity: Given intervals i, j, i', j' such that $Adj(i, j)$, $Adj(i', j')$, $Cong(i, i')$, $Cong(j, j')$, then $Cong(k, k')$, where $k = i+j$ and $k' = i'+j'$.

Now to guarantee the bi-infinitude of G, we adopt the following axiom:.

Theorem 4 *(Bi-Infinity of G) Let any interval i be given; then there exist exactly two intervals, j, k, such that $Cong(i, j)$ & $Cong(i, k)$ & $Adj(i, j)$ & $Adj(i, k)$ & one of j, k is left of i and the other is right of i.*

Proof. Given an interval i, by definition it is bounded, so there exists regions that are left of i and regions that are right of i. Assume a region m to (say) the right of i. (The case to the left is handled exactly analogously.) If $Adj(m, i)$, then, by the Translation axiom, there is a unique interval j such that $Cong(j, i)$ and j is left-end equivalent with m. If not-$Adj(m, i)$, then let f be the fusion of all intervals p such that $Betw(i, p, m)$. f is an interval. Then, by Translation, there is a unique interval j with $Cong(i, j)$ and j left-end equivalent to f. Combinig this with the analogous argument for the case to the left of i completes the proof. ■

Since bi-extension obviously iterates, this already insures that G is "bi-infinite" in the sense of containing as part the fusion of the minimal closure of any interval i under the operation of "*bi-extension*" defined in the theorem. (This closure is proved to exist in Lemma 3, below.) But we can do better and also insure that G is exhausted by iterating the process of flanking a given interval by two congruent ones as in Bi-infinity. This is just the Archimedean property, derived below. Toward this end, call an interval l an (*immediate*) *bi-extension* of interval i—$BiExt(l, i)$, or $biext(i) = l$—just in case $l = j+i+k$, where j, i, k behave as in the Bi-infinity theorem.

Lemma 2 *Let i* and *j be intervals such that $i < j$; then $\neg Cong(i, j)$.*

Proof. For a contradiction, assume $Cong(i, j)$. There are three possible cases: (1) i is left-end equivalent to j; (2) i is right-end equiv. to j; (3) i is neither. Cases (1) and (2) are argued in exactly the same way. For definiteness, assume case (1). By Bi-infinity, there exists i' extending i to the left with $Cong(i', i)$ and $Adj(i', i)$, hence $Adj(i', j)$. But then, by the hypothesis for reductio and transitivity of $Cong$, it follows that both i and j qualify as the unique right extension of i', as required by Bi-infinity, and since, by hypothesis of the Lemma, $i \neq j$, this is a contradiction. Case (2) is argued exactly analogously, considering i' as extending i to the right.

In case (3), let k_L be the fusion of all parts x of j such that $L(x,i)$ and let k_R be the fusion of all parts x of j such that $R(x,i)$. k_L and k_R are intervals. (Easy exercise.) Clearly, $k_L + i + k_R = j$ and this sum is discrete (all three pairs discrete). Now, let j' be (say) the right extension of j, i.e. $Cong(j',j)$ and $Adj(j',j)$. By Translation, let i' satisfy $Cong(i',i)$ with i' left-end equivalent to j'.By Translation again, let k'_L satisfy $Cong(k_L,k'_L)$ and $Adj(i',k'_L)$ with $L(i',k'_L)$; and let k'_R satisfy $Cong(k_R,k'_R)$ and $Adj(k'_L,k'_R)$ with $L(k'_L,k'_R)$. Then by Additivity, $Cong(j,i'+k'_L+k'_R)$, so, by the uniqueness of (right) extension of j as required by Bi-infinity, we have $j' = i'+k'_L+k'_R$, whence $i' < j'$, whence $i' \neq j'$, but then both j' and i' qualify as the unique right extension of j, a contradiction. ■

Now we can characterize G. Toward that, let X be any class of intervals such that an arbitrary but fixed interval $i \leq G$ is one of the X and such that if $k = biext(j)$ for j any of the intervals of X, then k is also in X. Call such X a "closure of i under *biext*".

Lemma 3 By axiom 2, there is an individual which is the common part of the fusions of each class X which is a closure of i under *biext*, which we call their *meet* or *the minimal closure i* of i under biext*. (Since i is stipulated to belong to any such X, the meet is non-null, as required in mereology.)

Proof. Immediate from axiom 2. ■

Given a fixed "unit" interval, i, we define the "right-half" or "postive half" i^+ of i^* as the fusion of i and all intervals j such that $R(j,i)$. Then we define the "left-half" or "negative half" of i as the fusion of all intervals j such that $L(j,i)$.

By the criterion for identity of mereological objects, the meet i^* of Lemma 3 is unique. We now can prove a theorem characterizing G as this meet:

Theorem 5 *(Characterization of G): Let G be the fusion of the objects in the range of the quantifiers of our axioms; then G = i*, the fusion of the minimal closure of i under biext.*

Proof. Suppose, for a contradiction, that $G \neq i^*$. Since, by stipulation, $i \leq G$ and G is closed under *biext*, we have that $i^* < G$. Then some part p, indeed (by axiom 5) an interval $k \leq G$ satsifies $\forall j[Int(j) \wedge j \leq i^* \rightarrow L(k,j)] \vee \forall j[Int(j \wedge j \leq i^* \rightarrow R(k,j]$; therefore, by definition, $L(k,i^*) \vee R(k,i^*)$. Let's suppose it's $R(k,i^*)$. (The other case is argued exactly analogously.) Let 'i^+' designate the positive or right half of i^*. Clearly

i^+ is connected; and by our betweenness criterion, it is also "bounded", so an interval. Therefore, by the Translation axiom, there is a unique interval $m \leq i^+$ with the properties (1) m is right-end-equivalent to i^+, and (2) $Cong(m, i)$. But this leads to contradiction, as follows: Note that i^+ consists of the fusion of class K satisfying (i) i is in K and (ii) if j is in K, then there is a unique j' adjacent and right of j with $Cong(j', i)$ and j' belongs to K, and (iii) for any other K' satisfying (i) and (ii), $K \subseteq K'$. Now regarding the relationship between m and K, there are four cases to consider: (a) All the K are to the left of m; (b) Adjacent to m on the left is one of the K; (c) One of the K—call it h—properly includes m, i.e. $m < h$; or (d) One of the K—call it h—overlaps m but neither is proper part of the other. Case (a) is ruled out since then i^+ then runs out before reaching m, contradicting that $m \leq \iota^+$. In case (b), it follows that m itself is one of the K; but then another, h', is in K, hence $h' \leq i^+$ with $Cong(h', i)$ and $R(h', m)$, but this is impossible because of property (1) of m above. In case (c), since both m and h are congruent to i, this contradicts Lemma 2 above that, if $j < k$, then $\neg Cong(j, k)$. (Also, then there would be h' in K and extending h to the right, contradicting that m is right-end equivalent to i^+.) Finally, in case(d), some h', in fact with $R(h', h)$ and $Adj(h', h)$, is in K, and $h' < m$ but both m and h' are congruent to i, again contradicting Lemma 2. Thus each of the four cases implies a contradiction, which shows that the assumption of such m, hence of a $k \leq G - i^*$, must be wrong. Thus, taking account of the exactly parallel argument for the left (negative) half of ι^* , it follows that $G = i^*$.[8] ■

Finally, we need a guarantee that any interval has a unique bisection. But that can now be proved as a theorem:

Theorem 6 *(Existence and uniqueness of bi-sections): Given any interval i, there exist intervals j, k such that $j < i$ & $k < i$ & $j|k$ & $j + k = i$ & $Cong(j, k)$; and j, k are unique with these properties.*

Proof. Let i be any interval. For any interval j, let j^{+r} be the fusion of j and the right bi-extension of j. So we need to find an interval j that is left-end-equivalent with i such that $i = j^{+r}$. Let k be any interval such that $k < i$. Without loss of generality, assume that k is

[8]Note that this result, expressing that G is Archimedean, is quite surprising as no axiom explicitly contains an "extremal clause" to the effect that the intervals of G are *only* those that are part of the fusion of those obtained by repeated applications of *biext* starting with a given interval. Nor do we have an induction axiom for properties of intervals, although, of course, in light of Theorem 3, such an induction principle could be derived from mathematical induction based on an N-structure. Alternatively, one could derive that from properties of minimal closures, à la Frege.

left-endequivalent with i. We have that $i - k$ exists, and is an interval. If k is congruent with $i - k$, we are done. So suppose not. Either $|k| < |i - k|$ or $|i - k| < |k|$. Let l be an interval that is congruent to the smaller of those two and is left-end-equivalent with i. So $l^{+r} \leq i$. (In fact $l^{+r} < i$). Now let j be the fusion of all intervals m such that m is left-end-equivalent with i and $m^{+r} \leq i$. Clearly $j^{+r} \leq i$. If $j^{+r} = i$, we are done. So suppose that $j^{+r} < i$. Let $n = i - j^{+r}$. Then n is an interval. Let p be an interval such that $p < n$ and, without loss of generality, suppose that p is left-end equivalent to n. Let q be $n - p$. Without loss of generality, assume that either $|p| = |q|$ or $|p| < |q|$. So $p^{+r} \leq n$. Let j' be the fusion of j and an interval congruent to p immediately on its right. (First let p_1 be congruent with p and right-end-equivalent with j. Then let p_2 be the right bi-extension of p_1.Then $j' = j + p_2$.) An application of Additivity shows that $j'^{+r} = j^{+r} + p^{+r}$, and we have $j^{+r} + p^{+r} \leq i$.This contradicts the definition of j as the fusion of all intervals m such that m is left-end-equivalent with i and $m^{+r} \leq i$.

For uniqueness, given interval i, suppose both $i = j_L + j_R$ with $Cong(j_L, j_R)$ and $Adj(j_L, j_R)$, and also $i = k_L + k_R$ with $Cong(k_L, k_R)$ and $Adj(k_L, k_R)$, with neither j_L nor $j_R = k_L$ or k_R. Suppose without loss that $k_L < j_L$(and so $j_R < k_R$). Let $m = j_L - k_L$. Then $j_R + m = k_R$, whence $Cong(j_R + m, k_L)$. Now let m' be congruent to m and adjacent to m to its right. Then $Cong(j_R - m', k_L)$. But we have $j_R - m' < j_R < j_R + m = k_R$ (where j_R is discrete from m). Since $Cong(j_R - m', k_R)$, this contradicts our lemma that if $i < j$, then $\neg Cong(i, j)$. ■

By repeated application of bi-sections, we can, in effect, approximate any locus along G to within any desired accuracy with sufficiently many nested intervals, whose least after k subdivisions is of norm 2^{-k} assuming the initial "unit interval" i is of norm 1. (Here we are speaking in our metalanguage, not yet having reconstructed the norm function in our object language.) One natural strategy that now suggests itself is to *define* an exact locus or "point" as a "Cauchy sequence" of such decreasing intervals. As a warm-up example, let us construct an endpoint—say the left—for a given arbitrary interval i. That will simply be the set of all subintervals j of i obtained by successive subdivisions into equal parts such that for all j, there is no $p < i$ such that $L(p, j)$. In point-based 1-dimensional geometry, if we arbitrarily set the left-endpoint of $i = 0$, this corresponds to the Cauchy sequence: $< \frac{1}{2^k} >$, $k = 1, 2, ...n...$, converging to 0. Indeed, we can introduce '0' in exactly this way: let i be an arbitrary but fixed interval, oriented as just described. Then $0 =^{df} \bigcap$ [all sets S containing i and containing the left half of any subinterval j of i such that $j \in S$]. Similarly, we could define 1, replacing 'left' (L) with 'right' (R). (Below, however, we give a definition in terms of Cauchy

sequences of intervals increasing to the right, in conformity to the rest of the positive half of G.) Notice that these "objects", whether thought of as numbers or as points, are not claimed to be parts of G. On the contrary, they are part of a *superstructure* that we construct over the mereological-interval structure of G.

In general, we define a sequence $< j_i >$ of intervals increasing to the right (or left, for negative reals) to be *Cauchy* just in case, for any interval, ε, there exists N such that for any $m > k > N$, $j_m - j_k$ is an interval, $R(j_m - j_k, j_k)$, and $|j_m - j_k| < |\varepsilon|$.(This last expression was defined contextually above. By assumption that $< j_i >$ is increasing to the right, $j_m - j_k$ exists and $R(j_m - j_k, j_k)$.Similarly for sequences increasing to the left. Note the role of axiom 5, guaranteeing that arbitrarily small intervals are values of ε.)

By repeated application of Bi-infinity and Translation, we can always avail ourselves of Cauchy interval sequences $s = < s_j >$ increasing to the right, beginning with our fixed unit interval, i, for positive reals (to the left, beginning with $-i$ for negative reals), i.e. such that $R(s_{k+1} - s_k, s_k)$ ($L(s_{k+1} - s_k, s_k)$); the fusion of all the intervals s_j forms an interval.[9] This proves convenient in giving a second representation of real numbers as intervals in G itself, which in turn—as we shall see below—augments the reach of reconstructions that don't rely on set theory. Thus, we will have available two relative interpretations of the classical continuum, $\mathbb{R}$, based on G: (1) equivalence classes of Cauchy sequences of intervals of G, or canonical ones from each class; and (2) fusions of canonical Cauchy interval sequences, as just indicated.

We'll return to (2) below. First, let us pursue (1) in some more detail. We want to construct the reals over the fixed interval i already associated with [0,1]. The first step is to identify the binary rationals as the appropriate subintervals left-endpoint-equivalent to i, obtained by iterated subdivisions licensed by the Bisection axiom. Thus, each rational of the form $\frac{n}{2^k}$, where $n = 1, 2, ..., 2^k - 1$, corresponds 1-1 with the left-endpoint-equivalent subintervals of i determined by the k'th stage of bisections. (The reader will have noted that the full binary tree of Baire space is in effect generated by these subdivisions.) The next step is to identify arbitrary reals in $(0, 1]$ with increasing Cauchy sequences of these subintervals.[10]

[9] Since we don't recognize a null interval, 0 is conventionally defined either as above or as a right-ward proceeding (nested) sequence starting with an interval, call it $-\frac{i}{2}$, congruent to the left half of i, choosing at each stage, k, the right half subinterval of $-\frac{i}{2^{k-1}}$.

[10] Of course, sequences may omit both left and right subintervals at a given stage, corresponding to a '0' in the binary numeral representation of the real in question.

The natural ordering of the binary subintervals of i implicit above is this: $j \prec k \leftrightarrow^{df} k - j$ exists (is non-null) & $R(k-j, j)$.

The next step is to extend this ordering to the increasing Cauchy sequences of intervals. We set $< r_i > \prec < s_i >$ just in case $\exists\varepsilon\exists N\forall k > N[\varepsilon$ an interval & $|s_k - r_k| > |\varepsilon|$ & $R(s_k - r_k, r_k)$ (Recall that the norm-notation was defined contextually using $Cong$, above.)

3 Recovering $\mathbb{R}$

The pieces are now in place to prove a first recovery theorem:

Theorem 7: *The ordered structure of binary intervals wthin i together with the Cauchy sequences of them is order-isomorphic to the classical real numbers of $(0, 1]$ (in their natural ordering, $<$).*

Proof. There is a 1-1 invertible map φ from the binary intervals of i to the binary rationals of $(0, 1]$. Define φ as follows: Set φ of the left interval, call it $\frac{i}{2}$, of i resulting from the 1st subdivision $= \frac{1}{2}$; after the *k'th* subdivision, set φ of the left-most $= \frac{1}{2^k}$, of the next left-most $= \frac{2}{2^k}$, ..., of the next to right-most $= \frac{2^k-1}{2^k}$. (φ of the right-most is of course always $= 1$.)[11] Clearly φ is order-preserving. To extend this to the increasing Cauchy interval sequences, map each such, of the form $< \ s_1, s_2, ..., s_n... >$ to the increasing rational Cauchy sequence determined by φ, viz. $< \varphi(s_1), \varphi(s_2), ..., \varphi(s_n), ... >$. Call this extension of φ φ'. That φ' is 1-1 and onto the increasing binary rational Cauchy sequences is immediate from the properties of φ. That φ' is order-preserving is also clear: if $< r_k > \prec < s_k >$, then beyond some N (given in the definition of this ordering, above), $\exists n$ such that the corresponding rational differences, $\varphi(s_m) - \varphi(r_m) > 2^{-n}$, for any $m > N$, which defines order on these rational sequences. ■

Now we can extend this to the whole positive half-line, $(0, \infty)$ by applying the same procedure to right-extensions of i by any number of intervals each congruent with i. E.g. we map the interval $i + j$, where $Cong(i, j)$ & $Adj(i, j)$ & $R(j, i)$ to $(0, 2]$ (appealing to the Bi-infinity theorem), iterating this procedure to cover all intervals of the form $(0, n]$. Thus, we have:

[11] Since, e.g., $\frac{2}{2^k} = \frac{1}{2^{k-1}}$, etc., it appears that φ is many-one; but really it isn't as the the result of proceeding stepwise to the right simply adds a congruent interval adjacent to the preceding, so that the result is an interval, and in the case of an even number of steps, it is always = an interval obtained at an earlier subdivision. E.g. in one rightward step at the kth subdivision, we get left-most-$\frac{i}{2^k}$+ next-left-most-$\frac{i}{2^k}$ = left-most-$\frac{i}{2^{k-1}}$, etc. (where the '+'here is mereological summation).

Corollary 1 *The theorem statement holds for all intervals of the form* $(0, n]$, *hence for the whole positive part* G^+ *of* G *as order-isomorphic to the positive reals,* $(0, \infty)$.

Proof. The only thing to check, in addition to what has already been established, is that the map just introduced—call it φ''—from G^+ to $(0, \infty)$ is indeed defined on all of G^+ in the sense that no part $p \leq G^+$ is discrete from all the intervals on which φ'' is defined. Suppose, to the contrary, there is some such part, p. Then p must be discrete from each binary rational interval left-end equivalent to i, and in particular discrete from any of the intervals obtained from i by any finite number of applications of the *biext* operation introduced above. But then p would be discrete from the fusion of all such intervals, which fusion = exactly the meet (= minimal closure of i under *biext*) that we proved = G in Theorem 5, above. This contradicts that $p \leq G^+$. ■

Corollary 2 *The theorem statement holds for all intervals of the form* $[-n, 0]$, *hence for the whole negative part* G^- *of* G *as order-isomorphic to the non-positive reals,* $(-\infty, 0]$.

Proof. Applying Translation and Bi-infinity, the constructions for the positive part of G can be shifted accordingly. (For the sake of the field operations, introduced below, it is simplest to reflect the positive intervals "about 0", i.e. proceeding leftward starting with $-i$ or subdivisions thereof.) The proof that the "mirror image" of the map φ'' is defined on all of G^- is exactly analogous to the proof of the first Corollary that φ'' is defined on all of G^+. ■

To extend the isomorphisms of the Corollaries to cover the whole of G simply stipulate that, for every interval j of the negative part of G and every interval k of the positive part, $j \prec k$.Thus we have established:

Theorem 8: *G (i.e. $(G, \prec)$) is order isomorphic to R (ordered by $<$).*

Call the isomorphism of the latter theorem Φ.

With respect to the field-algebraic structure of $\mathbb{R}$, we can proceed in either of two ways. (1) We could regard it as additional structure of $\mathbb{R}$, built up in the usual way from the Cauchy seqences defining the reals, not bothering about any additional structure of G; or (2) we can introduce operations of "addition" and "multiplication" of intervals of G and prove that the order-isomorphism Φ is also an algebraic isomorphism. (2) is more interesting so let us pursue it. To define an interval sum operation, call it $i \oplus j$, we can first define this for binary intervals,which we can already express as $\Phi^{-1}(q)$, for q a binary rational of $\mathbb{R}$, where such an interval is either left-end-equivalent to i, if $q > 0$, or right-end equivalent to $-i$, if $q < 0$, where $-i$ is the interval satisfying $Cong(i, -i)$

& $Adj(i,-i)$ & $R(i,-i)$.Then $\oplus$ will simply be vector-addition along G. $\oplus$ is then extended to all of $\mathbb{R}$ by applying it "pointwise" to inverse images under Φ of the (binary) rational Cauchy sequences defining the reals in queston. To obtain "multiplication" of intervals, $i \otimes j$, either of two method may be used. Remaining entirely within G, we first define this product for inverse images of binary rationals as iterated interval-addition, $\oplus$, treating binary fractions of intervals in accordance with the distributive law of multiplication over addition.[12] Then $\otimes$ is extended to all of $\mathbb{R}$ by applying it "pointwise" to inverse images under Φ of the rational Cauchy sequences defining the reals in question. Alternatively, we can adapt the establshed Euclidean geometric method of introducing product of two lengths by working in the Euclidean plane. Here, of course, we work in $G \times G$, diagrammatically representing one copy of G as our abcissa, the other as ordinate, such that the left end of our unit, i, is the origin, where the two axes cross.[13] Then the product $j \otimes k$ is obtained by taking j as left-end equvalent to i along the abcissa, k as extending from the origin along the ordinate (either "up" if we are operating on $+k$, or "down" if we are operating on $-k$). Next we construct the "hypotenuse" segment σ connecting the right-end of i with the "top"-end of k (if we're consider $j \otimes +k$, "bottom"-end of k if it's $j \otimes -k$). The value of the product is then represented as the interval m along the ordinate from the origin to where the segment, call in σ', meets the ordinate, where σ' extends from the right-end of j and lies parallel to σ. (σ' forms a second "hypotenuse", so we have two similar triangles. Then the definition of product derives directly from the fact that $\frac{|m|}{|k|} = \frac{|j|}{1}$.)[14] Now the definition of the operation $\otimes$ is extended to all of $\mathbb{R}$ via rational Cauchy sequences, as in the first method.

Recalling that $\Phi^{-1}(0)$ was introduced via a rational-interval Cauchy sequence, one checks that the algebraic laws of the field operations are preserved under Φ. Thus, we have

Theorem 9: *G $(G, \prec, \oplus, \otimes)$ is ordered-field isomorphic to R $(R, <, +, \cdot)$.*

[12] E.g., $\Phi^{-1}(2.5) \otimes \Phi^{-1}(3.25)$ is computed by vectorially adding twice the (unit) interval i and half-i, call this $2.5i$, and then vectorially tripling this and adding a quarter of $2.5i$ to obtain the answer.

[13] Justifying this method on the present pointless basis would require adopting some further primitives and axioms to extend our methods to the Euclidean plane. We would need further equivalence relations of "end-equivalence" to replace reference to "the point where two non-collinear intervals meet"; and we would need a relation of angle-congruence for purposes of constructing paralells. All this will be carried out in further work on two-dimensional continua.

[14] Note that we're now in a position to define the norm, $|j|$, for any interval j, based on the isomorphism Φ.

Although these isomorphism theorems rely on the characterization theorem (5) of G, which already expresses that G is Archimedean, it is worth pointing out that Theorem 9 implies that G satisfies a statement of the Archimedean property closer to the one commonly given for $\mathbb{R}$. Here is one way to formulate this. As will be suggested below, assume as part of our mereological universe a natural-numbers-structure, N, constituted entirely of atoms, which we designate $0_N, 1_N, 2_N, ...$[15] Then we can speak of functions from N $('', +_N, \bullet_N)$ to intervals of G by quantifying plurally over (unordered) pairs, $(n, j) =^{df} n + j$ (where this is nominalistic summing). Now inductively define a map $\varphi : N \rightarrow$ Intervals of G via

(i) $\varphi(0_N) = i$

(ii) $\varphi(n_N +_N 1_N) = biext(\varphi(n_N))$.

Now define 'i divides k (mod 3)' to mean: $k = \varphi(m_N)$, some m_N. Then we say G *is Archimedean* just in case

$$\forall j \exists h[j \text{ an interval } \rightarrow h \text{ an interval } \& \ i \text{ divides } h \text{ (mod } 3) \ \& \ j \leq h].$$

Thus we have

Corollary G is Archimedean.

Proof. This follows from the Characterization Theorem on G as the minimal closure of i under *biext*. For a contradiction, suppose that some interval $j \leq G$ is not covered by any interval h obtained by iterating *biext*, starting with i, any finite number of times. Then, by construction of G, for some $j' \leq j$, $j'|G$, so $j \not< G$, contradiction. (Cf. the proofs of Corollaries 1 and 2, via the definedness of the maps described there.) ■

Finally, we would like to be assured that the above development of $\mathbb{R}$ over G is independent of the starting interval, i, i.e. that starting with any other, j, leads to essentially the same recovery theorems, and even that the minimal closure of j under *biextension* is indeed G itself. That can be arrived at as follows.

Let j be any interval of G other than i.Now minimally close j under *biextension*, calling the result j*. Then we proceed in two steps: (1) We can carry out the whole of the above construction to produce an ordered-field isomorphism Φ' from the binary intervals and Cauchy sequences thereof between G' and $\mathbb{R}$,where G' and Φ' are introduced just as G and Φ were but substituting reference to j for that to i throughout. (Think of a transformation of the interval structure of G based on i combining a suitable translation and either a shrinking, in case $|j| < |i|$, or a

[15] This can itself be carried out along lines of Hellman (1996) but ignoring modal operators for present purposes.

stretching, in case $|i| < |j|$, or neither, if $Cong(i, j)$.) Step (2) consists of demonstrating that, in light of (1), the point-free continuum $G' = j^*$ based on j is indeed $= G$. That makes essential use of the Archimedean property of G', afforded by carrying out the proof of the last Corollary with G' in place of G, *mutatis mutandis.*

Theorem 10: *The point-free continuum $G' = j^*$ based on interval j ($\leq G$ and $\leq G'$) $= G$.*

Proof. By definition of G as minimal closure of i under *biextension*, $G' \leq G$. For since, by hypothesis, $j \leq G$, we have that for some finite n, $j \leq biext^n(i)$, i.e. j is part of the result of the n'th iterate of *biext* applied to i. Then it is straightforward to show that any k covered by finitely many applications of *biext* starting with j is also covered by finitely many applications of *biext* starting with i.We need to show the converse, that $G \leq G'$. First we claim that $i \leq G'$. This follows from the Archimedean condition displayed above, applied to G', interchanging the roles of 'i' and 'j' to produce an h ($\leq G'$) such that $i \leq h$. Next we argue by induction that any interval k obtained from i by repeated application of the *biextension* operation will be accessible from j in the same sense as i is, i.e. by appealing again to the Archimedean condition above for G', substituting 'k' for 'j' and 'j' for 'i'. Then, from the definition of G, since every interval m of G is covered by the result of some finite number of iterations of *biextension* based on i, it follows from the induction that $m \leq G'$, as well, whence $G \leq G'$, and therefore $G' = G$. ■

4 Topological Models

We now present two topological models for our axiomatization. These illustrate some of the Aristotelian notions of contiguity and continuity.[16] The exercise will also serve to remove any lingering doubts concerning the consistency of our axiomatization, if there are any.

For both models, the background meta-theory is the ordinary, Dedekind-Cantor account of the real numbers, with their usual topological properties. An open set S of real numbers is said to be *regular* if S is identical to the interior of its closure.[17]

Define a real number r to be an *interior boundary* of a set S, if $r \notin S$, but there are numbers s, t such that $s < r < t$ and the open set

[16] For a discussiion and comparison of these concepts, see Hellman and Shapiro [2012].

[17] Cartwright [1975] argues that in 3-space, all and only regular open sets are "receptacles", regions of space that physical objects can occupy. Cartwright's theory is at least partly Aristotelian.

$(s,t) - \{r\}$ is a subset of S. So, for example, the number 1 is an interior boundary of the union of $(0,1)$ and $(1,2)$. Regular open sets are those open sets that have no interior boundaries.

The domain of our first model consists of all non-empty, regular open sets of real numbers. The parthood relation is just the subset relation, as might be expected. Let Π be a non-empty set (or plurality) of non-empty, regular sets. Define $SUM(\Pi)$ to be the interior of the closure of the union of Π. This is the fusion relation. So to get the fusion of a set of regions in our model, first take the union of the sets, then the closure of the result, and then the interior of that. The result is, again, a regular, open set. And it is straightforward to verify that our Axiom 2, of fusion (or whole comprehension) is satisfied:

,

$$(\forall ww)(\exists x)(\forall y)[y \circ x \leftrightarrow (\exists z)(z \ \eta \ ww \wedge z \circ y)]$$

.

This model nicely recapitulates some of Aristotle's account of continuous objects. Consider, again, the intervals $(0,1)$ and $(1,2)$. Those are "contiguous", since there is nothing of the same kind in between them. Indeed, there are no members of our domain in between. The only thing "between" them is the real number 1, and $\{1\}$ is, of course, not a regular open set. These two intervals are also "continuous", in Aristotle's sense, since when we put them together—when we take their sum—we obtain the interval $(0,2)$. That is, the "boundary" disappears and they become a single interval, a unity.

The proper definitions of the other primitives in our axiomatization are as straightforward as can be. Recall that $L(x,y)$ intuitively means that the region x is entirely to the left of y. Let X and Y be non-empty, regular open sets of real numbers. Then define X to be $LEFT$ of Y just in case every member of X is smaller than every member of Y. It is trivial to verify that the relevant axioms are satisfied.

Recall that an "interval" is defined to be a connected, bounded region. In the model, the "intervals" are just the open intervals, (a,b), with $a < b$. And, of course, the Gunkiness axiom 5 is also trivial:

$$(\forall x)(\exists j)(Int(j) \ \wedge \ j < x).$$

The notion of "left-end-equivalence" is also straightforward. Two intervals are left-end-equivalent just in case they have the same left endpoint. And, of course, "congruence" of regions is defined to be congruence of regular, open sets. Verifying the remaining axioms is also trivial.

Our second topological model is a sort of dual to the first, as its domain consists of certain *closed* sets of real numbers. It is, in one sense, a little more Aristotelian, since it does allow that intervals have endpoints—although these endpoints are not regions in the domain.

Say that a set S of real numbers is *regular closed* if S is identical to the closure of its interior. A real number r is an isolated point of a set B if $r \in B$, and there are numbers s, t such that $s < r < t$ and the open set $(s, t) - \{r\}$ is disjoint from B. So, for example, 1 is an isolated point of $\{1\}$, and also of the union of $\{1\}$ with $[2, 3]$. Regular closed sets have no isolated points.

The domain of our second model is the set of non-empty, regular closed sets of real numbers. As with our first model, the parthood relation is the subset relation. If Π is a non-empty set of non-empty regular closed sets, define the SUM of Π to be the closure of union of Π. It is straightforward to verify that this SUM is regular closed, and that Axiom 2 of fusion is satisfied.

Perhaps this model better captures the Aristotelian notion of continuous objects. The sum of $[0, 1]$ and $[1, 2]$ is, of course, $[0, 2]$. Here the boundary point(s) of the contiguous intervals is "absorbed" into the sum.

With this model, we must be a little more careful when characterizing some of the relations in our theory, even the defined ones. Consider, for example, the two closed intervals $[0, 1]$ and $[1, 2]$, both of which are in our model. As sets, those are not disjoint, as they have a member, 1, in common. Recall, however, that in our definition of overlap:

$$x \circ y \leftrightarrow (\exists z)(z \leq x \ \wedge \ z \leq y),$$

the quantifier ranges over the *regions* in the model. So $[0, 1]$ does not overlap $[1, 2]$, since there is no member of the domain—no regular closed set—that is a part of (i.e., a subset of) both. The regions are indeed discrete. In our second model, "intervals"—bounded and connected regions—are closed intervals $[a, b]$.

Let A be the union of $[0, 1], [2, 3], [4, 5],. \ . \ .;$ and let B be the union of $[1, 2], [3, 4], [5, 6], \ . \ . \ .$ Then A does not overlap B—they are discrete—even though their intersection, as sets, is infinite.

Our definition of "Left" is similarly nuanced. If A and B are regular closed sets, then define A to be $LEFT$ of B just in case for every r in A and every s in B, either $r < s$, or both $r = s$ and r is a boundary of both. So $[0, 1]$ is to the left of $[1, 2]$. It is trivial to verify that the axioms of our theory are all satisfied in this second model.

This second model does have some rather strange or, at least intuitively un-Aristotelian regions. Consider the union of the closed

intervals $[.1, .9]$, $[.01, .09]$, $[.001, .009]$, . . . and $\{0\}$. This set is regular closed, and so is a region in our model. However, it has an actual infinity of discrete parts and a sort of loose point $\{0\}$ at its left end.

5 Interlude: A Brief Dialogue

Objection: If this is really coherent, it would indeed reënforce that "non-punctiform" and "indecomposable" are very different attributes as applied to continua.[18] *But the connection between them may be greater than the above seems to imply. Indeed, one may question whether it really makes sense to say of G that it has any proper parts at all! Let's take your interval i: You suppose that it makes sense to speak of i and its "negate", Neg(i) = G − i. But consider the question of the two places where i "meets" Neg(i): Do they touch there? Surely there must be a point at each of those places, either as part of i itself or of Neg(i). But of course, as your system allows, i can be shifted anywhere along G so that any place can serve as a boundary of an interval, in which case G is composed of points after all.*

Reply: It is indeed part of the standard punctiform conception that boundaries of (bounded) intervals must exist, and then they must be part of one or the other of an interval and its complement.[19] But the above theory of G makes sense on the basis of a *different* conception. According to it, let's consider your question, whether i and Neg(i) "touch". Well, we don't make sense of that, except to say that, if you mean, "Is there anything in between i and Neg(i)?", the answer is clearly no; and this is perfectly compatible with the two parts not overlapping. After all, a very similar thing happens in the case of your Dedekind cuts in the rationals used to define irrational numbers: the lower and upper sections are disjoint (the set-theoretic analogue of our "discrete"), yet there is no "boundary" in the sense of an lub of the lower section or a glb of the upper, until *by fiat* one is introduced by defining it to be the cut (or a section thereof) itself! We all know and love this move by Dedekind, and we know how well it works. But that move does not establish that a corresponding point really exists ("was there all along") on any actual

[18] In the opposite direction, the intuitionistic continuum is "indecomposable" but, in some sense of 'point', is entirely composed of points. Although not all pairs of reals are orderable (so that "exact location" cannot be attributed), still they are specfied with infinite precision relative to the everywhere dense rationals.

[19] It seems also to have been Aristotle's conception that "breaking" a line segment results in (or "actualizes") boundary points, endpoints of two new sbintervals. This may in fact have been Aristotle's notion of "indecomposability". In that case, note how different it is from the intuitionistic conception. There, breaking a continuum results in a *loss* (some "syrup sticking to the knife", as it were), whereas, on Aristotle's conception, the result is an *addition* (of endpoints) to the structure.

1-dimensional continuum (if there are any), much less on any that can be coherently conceived. And it is simply not a move that we are forced to make, either on conceptual or on practical grounds. Instead, we can, if we like, carry out a Dedekind style construction as a *superstructure* over G, preserving all the advantages of classical analysis while resisting the attribution of a point-ontology.

Objection: The analogy with Dedekind cuts (for irrationals) is flawed: Your G is supposed to model a continuum, so the analogue of a division of G into two discrete yet adjacent parts would be a Dedekind cut in the reals, not in the rationals. But the Dedekind continuum is complete, so there are no cuts in the reals without the corresponding real belonging to one segment or the other. This shows, does it not, that e.g. your Bisection axiom really makes no sense unless a point is added at the place where the two parts meet.

Reply: Agreed, the analogy is imperfect; still, imperfect analogies can be of heuristic value. But it doesn't follow that points need to be added "where [the] two parts meet", for, as we said above, "where they meet" is language belonging to the Cantor-Dedekind theories of a pointful continuum, but it is foreign language that can't really be translated into the theory of G. Point-like "places" simply are not recognized; *that* is the point! Yes, they can be introduced as superstructure, as shown above, but that doesn't require revising the description of G, or thinking that "really" the points are "there" on G. Just as we can say that two people believe in the same god(s) without thereby implying that there exist god(s) in which they both believe, so we can speak (as we have spoken) of two intervals' being, say, "left-end equivalent" without thereby implying that there is a special further entity called "the left end (as a point)" shared by both segments. Even in cases where abstraction based on an equivalence relation may be reasonable and useful, e.g. as when we speak of "income levels" based on persons' or families' "sharing similar enough incomes", that doesn't mean—at least, without much further argument—that we have to recognize "income levels" as entities or that it is somehow incoherent not to.

6 Non-set-theoretic versions

There were two main places in the above reconstruction where set theory seemed to play an indisepnsable role: (1) in the use of Cauchy sequences of binary intervals; and (2) in the use of isomorphisms between G and $\mathbb{R}$. In the latter case, appeal to set theory is, in a sense, not really problematic with regard to the autonomy of the interval structure of G, since, after all, we're reasoning about the relation of that structure to that of the classical continuum, which is fundamentally set-theoretic. The first

case, however, is problematic, as the autonomy of G is threatened by reliance on enough set theory to build up $\mathbb{R}$, with its ontology of points, etc. Is there an alternative way of introducing interval sequences that improves on this?

Indeed there is. All that is needed is to expand the theory of G by specifying that, additional to G but entirely discrete from it, there is a denumerable infinity of mereological atoms that form a natural-numbers structure ("N-structure"). That is readily expressed with mereology and plural quantification. (Cf. e.g. Hellman [8].) Now the role of ordered pairs of the form $< n, j >$, where n is an atom of the N-structure and j is an binary rational interval of G, can be served by the (unordered) mereological sum, $n + j$.[20] Plurally quantifying over such things then enables *plurally quantifying* over binary rational interval sequences. Those that satisfy the Cauchy convergence condition serve as our reals. The upshot is that, now, we bypass set theory entirely in the development of our G-interval structure, *even to the extent of being able to carry out a great deal of classical analysis without even recognizing real numbers as objects!* Instead of speaking of reals as individual objects, we quantify plurally over the increasing binary rational Cauchy interval sequences corresponding to them (as their limit points, on standard classical theory). This fulfills at least the non-punctiform aspect of the Aristotelian conception of contiua in a clearly non-parasitic way. Points aren't introduced at all; rather, their mathematical roles are performed by surrogates within the thoroughly non-punctiform framework of G. We can even claim to capture Aristotle's idea that "points exist only *potentially*", since, as already seen, we *can* introduce real numbers or points as objects serving as the limits of our converging interval sequences, but that is not forced on us.

But, in order to make good on this last claim, it needs to be checked that at least a good portion of classical analysis can indeed be constructed within the G-framework (with plural quantification but no set theory, as just described). In particular, how are we to reconstruct quantification over functions from reals to reals if we can only plurally quantify over Cauchy interval sequences. This is no problem so long as our functions are continuous (or at least continuous on a co-countable domain of reals). The key here is that continuous functions are determined

[20] Although the intervals of G already allow encoding of natural numbers and the arithmetic operations, representation of sequences of intervals and other functions within a nominalist framework becomes cumbersome at best, since a stock of intervals would be serving two roles at once, that of natural numbers and that of items of sequences. Such problems are bypassed by positing an atomic N-structure discrete from G.

by their behavior at rational arguments. Now binary rationals are available directly as intervals. Suppose we have a sequence of all these representing rationals in the domain of definition of f, i.e. we refer plurally to countably many individuals of the form $< n, j >$, as defined above. A continuous function f assigns each of these a real value, represented over G as a Cauchy interval sequence, $f(< n, j >) = f_1^{n+j}, f_2^{n+j}, ...$Now we would like to code all these (countably many) Cauchy interval sequences as a single interval sequence (which itself, of course, need not be convergent—it merely serves to encode the behavior of f at rational arguments, enumerated in an arbitrary fashion). One convenient way is to work with just the atoms n of the $< n, j >$ serving in the enumeration of rational intervals, coding the (set-theoretic) ordered-pair $< n, f(< n, j >) >$ as the sequence $n + f_1^{n+j}, n + f_2^{n+j}, ...$Then the restriction $f \restriction Q$ of f to rationals can be coded up in a single sequence by a dove-tailing construction on the countably many sequences of this latter form.[21] In this way, we bring continuous and co-countably continuous functions within the purview of plural quantification of the language of the G-interval structure.

With one more reductive step, we can improve further on this to bring also the isomorphism recovery theorems within the purview of the theory of G (or its expanded version, providing for an atomic N-structure). Instead of recovering real numbers as Cauchy interval sequences, we can avail ourselves of the second relative interpretation of $\mathbb{R}$ over G referred to above, toward the end of section 2, viz representing reals as certain intervals of G, viz. fusions of the right-(or left-) increasing Cauchy interval sequences introduced above. Here is how this helps in representing the isomorphisms of the recovery theorems without set-theoretic machinery.

As in the non-set-theoretic reconstruction just sketched, we reconstruct interval sequences via an atomic N-structure. But now instead of treating quantification over reals as plural quantification over the wholes, $n + j$, coding the (canonical) Cauchy interval sequences defining them, instead we quantify singularly over the fusions of the intervals making up such a sequence, which, as noted, are themselves intervals of G.

Before we can introduce functions from such "real intervals" of G (as we may now call them) to reals of the classical continuum, we need (or at least prefer) to identify a suitable target structure that is itself not essentially set-theoretic. Here we can appeal to arithmetic methods applied to our given N-structure. Coding signed integers $+m$ $(-m)$ conventionally as certain pairs of naturals, e.g. as $2^1 \cdot 3^m$ $(2^2 \cdot 3^m)$, and then

[21] This is modeled on the way countably many real numbers can be coded as a single real number.

rationals as (reduced) pairs of signed integers, we then have canonical Cauchy sequences $< q_j >$ of rationals coded as number-theoretic ordered pairs of the form $< j, \overline{q}_j >$, where $\overline{q}_j$ is the number coding the rational q_j.[22] Any two distinct such sequences differ at some place; thus, the fusion of any such sequence is just a uniquely determined whole of atoms of our original N-structure, with different sequences yielding different fusions. These fusions then can serve as our classical reals (which we call "reals over $\mathbb{N}$"). It is tedious, but routine, to define linear ordering and the field operations on these.

We still need a way of representing or coding mappings between real-intervals of G and fusions over the N-structure serving as classical reals . First, how can we code an ordered pair of the form $< k, \sigma >$, where k is a real-interval of G and σ is a (nominalistic) sum of "naturals" (atoms of the N-structure) coding a canonical Cauchy sequence of rationals representing a real? Well, if we stipulate that the mapping goes in the direction $k \rightarrowtail \sigma$, we can make do with an *un*ordered pair. But that can simply be taken as the fusion of k and σ, as these are always non-overlapping, as they are from different structures, one atomless, the other atomic. Finally, by quantifying plurally over fusions of this form $(k + \sigma)$, we achieve the effect of quantifying over mappings or functions from our G-structure to the reals over $\mathbb{N}$. In this manner, we can reconstruct the isomorphisms of the recovery theorems and prove their required properties within our theory of G together with an atomic N-structure, without ever using set-membership or quantifying over sets.

Finally, there is a readily available way of further reinforcing the idea that real numbers as "points" "exist only potentially": the above extension of the universe of G by an N-structure and the whole subsequent development of an $\mathbb{R}$-structure over it can be carried out under the supposition that such an extension is merely *logically possible.*[23] Indeed, even the theory of G itself can be carried out relative to the assumption that such a mereological-interval structure is merely a logical possibility. Even so, extensions of G by further structures, e.g. an atomic N-structure as above, are taken as *further* possibilities, relative to any given hypothetical G-structure, so that the "merely possible" status of real numbers as points is still recognized even within a thorough-going modal-structural treatment.

[22] Here, of course, j is a numerical index referring to an atom of our N-structure, not an interval of G.

[23] Thus, the above development of an extension of G by an atomic N-structure, etc., can be set in the modal-structural framework of Hellman [7], as improved in [8] and subsequent presentations. We do not claim that this captures what Aristotle meant by "potentially infinite".

7 Comparisons with Some Other Constructions

Without attempting anything like complete coverage of alternatives already in the literature, we present some comparisons that we hope will be instructive.

As our title states, we are concentrating on point-free constructions of the classical continuum, $\mathbb{R}$, and in further work we will develop a recovery of $\mathbb{R} \times \mathbb{R}$, which extends to higher dimensions, without primitives for any objects of lower dimension. In addition to forming the basis of classical functional analysis, these systems can be enriched to study geometric spaces of various sorts. So we are clearly working in the area (category) of *metric spaces*, not purely topological spaces. There have, indeed, been a number of efforts to develop topological spaces based on axioms governing *regions*, rather than sets of points, going back at least to work of Karl Menger.[24] Our own work only touches indirectly on these precedents, insofar as well-known topological spaces can be constructed from $\mathbb{R}$, $\mathbb{R}^2$,.and higher dimensions. But those are derivative from the metrical structure of these spaces, and so aren't purely topological.

Focusing, then, on point-free geometries, there is the noteworthy reconstruction of three-dimensional Euclidean geometry by Tarski [14]. This explicitly uses (atomless) mereology, to which is added the single primitive '*sphere*'. (In addition, the construction uses some set theory, e.g. in the key definition of *point* as the set of all spheres *concentric with* a given sphere, and also in recovering definitions of various kinds of sets of points, e.g. "*regular open*", etc.) A few clever definitions introduce the notions "*sphere A is externally tangent to sphere B*", "*sphere A is internally tangent to sphere B*", "*spheres A and B are externally diametrical to sphere C*", "*spheres A and B are internally diametrical to sphere C*", leading finally to "*sphere A is concentric with sphere B*": where $A \neq B$, and where, say A, is proper part of B, given two spheres, X, Y, externally diametrical to A and internally tangent to B, X and Y are also internally diametrical to B. Then '*point*' is introduced as already explained, and then the notion of two points being *equidistant from* a third is defined.

What about axioms? As his *Postulate I*, Tarski stipulates that "The notion of *point* and that of *equidistant from* satisfy all the postulates of ordinary Euclidean geometry of three dimensions." Then follow three "auxiliary postulates" governing *solids*, connecting *solid* and *part* with *regular open set* and *inclusion*. He then writes: "The postulate system given above is far from being simple and elegant; it seems very likely that

[24] See, e.g. Menger [11]. For further references and recent developments, see Roeper [12] and Gruszczyński and Pietruszczak [6].

this postulate system can be essentially simplified by using intrinsic properties of the geometry of solids." For the one-dimnesional case, that is precisely what our system above accomplishes. We may call an approach such as our own, in which no axioms govening even *defined* "points" are listed (or, in higher dimensional cases, governing defined concepts for any lower-dimensional objects), one of "honest toil". When it comes to two-dimensions, we will present a recovery of the key Archimedean property meriting the honorific "honest toil". Another departs from this but only fairly modestly; we can describe this as an instance of "petty theft" (first offiense). In comparison, Tarski's method seems an instance of "grand larceny". As his self-critical remark suggests, it is one thing to show that key primitive notions can be adequately defined by terms designating objects and relations of a given level (dimension); it is quite another to achieve a full-scale *reduction* of a point-based theory by *deriving* translations of its axioms induced by the definitions as *theorems* from axioms governing concepts pertaining entirely to the given level. In fairness, it should be mentioned that, as Tarski states, his postulates can be proved categorical, and they can be proved mutually relatively interpretable with standard point-based Euclidean geometry. Nevertheless, the achievement of a full-scale reduction is clearly more desirable, not only for the unitiy it achieves, but for establishing the autonomy and sufficiency of the conceptual machinery operating at a given level or dimension.

Much closer in content and method to the reconstructions of continua presented here is the recent work of Roeper [13], on what he calls "the Aristotelian continuum". Like the system presented above, his system charcacterizes a linear continuum on a point-free basis in terms of a structure of regions and intervals, and it is shown categorical and isomorphic to a classical, point-based continuum. The following are the main points of comparison:

(1) Roeper's axioms describe a continuum as a certain kind of region-based topological space, connected, locally connected, second-countable, also linearly ordered, complete, and separable. Our axioms do this but also describe the metrical structure of an ordered field. Thus, we have *congruence* of intervals as one of our primitives, whereas Roeper's system omits this.

(2) As part of his logic, in place of mereology and logic of plurals, Roeper uses a (first- and second-order) logic of mass terms. Thus, where we would express, e.g., "region r is entirely to the left of region s" as "any part of r is left of every part of s", Roeper's language expresses "r is everywhere left of s (everywhere)". Probably the two languages, as used in these respective reconstructions of continua, that is, in the presence

of other primitives and axioms governing them, are inter-translatable.

(3) Roeper draws heavily on region-based topology, a theory with two non-logical primitives, *limited* (that is, bounded or "finite in extent") and α *is connected with* β, written $\alpha \infty \beta$, intuitively meaning either α and β overlap or they "abut one another", along with thirteen axioms governing these, including *Coherence* (if $\gamma = \alpha \cup \beta$, then $\alpha \infty \beta$), *Continuity* (a topological version of a least-upper-bound or greatest-lower-bound principle), and *Countable Convex Cover* (akin to Separability). It should be emphasized that these topological axioms are not axioms of the Aristotelian continuum, but are instead (required and shown to be) derivable therefrom. However, the two primitives just listed are taken over in axioms 6, 7, and 8 of the latter system. Axiom 6 is essentially our definition of *bounded*; axiom 7 is a version of *continuity*; and axiom 8 is a version of *separability.* For purposes of describing the classical linear continuum, we define the two primitives, *limited* and *connected with*: for α *is limited,* see the definition of "*bounded*" above, sec. 2; for *connected*, we defined *Adj(r,s)*, "r, s are adjacent", and of course we have $r \circ s$, from mereology. Moreover, the crucial properties of *coherence*, *continuity*, and *separability* are not taken as axioms of our system, but rather proved as theorems, where for the first two properties, we use the plurals comprehension scheme of mereology (describing minimal closures), and for the third we use the representation of rational-length intervals via the theorem of bisectability. Regarding the important Archimedean property, Roeper's system can derive this from its version of continuity or completeness (as standardly done in point-based frameworks, e.g. à la Dedekind), whereas we prove this directly from our comprehension axioms and Translation.

Indeed, we can claim more: all Roeper's axioms are in fact derivable as theorems in our system, under the translation just given of his primitives, *limited* and *connected*. Thus we have a nice unification of systems, our geometry-*cum*-classical analysis and Roeper's regions-based topology-*cum*-ordered continuum. In sum, "honest toil" has paid off.

References

[1] Aristotle *Physics Book VI, The Basic Works of Aristotle*, R. McKeon, ed. (Randon House, 1941), pp. 316, ff.

[2] Bell, J.L. *A Primer of Infinitesimal Analysis* (Cambridge University Press, 1998).

[3] Bell, J.L. "The Continuum in Smooth Infinitesimal Analysis", P. Shuster, U. Berger, and H. Osswald, eds. *Reuniting the Antipodes: Constructive and Nonstandard Views of the Continuum* (Kluwer, 2001), pp. 19-24.

[4] Bishop, E. *Foundations of Constructive Analysis* (McGraw, 1967).
[5] Cartwright, R., "Scattered objects", in *Analysis and Metaphysics*, edited by Keith Lehrer (Dordrecht: Reidel, 1975), 153–171.
[6] Gruszczyński, R. and Pietruszczak, A. "Space, Points, and Mereology: On foundations of point-free Euclidean geometry", *Logic and Logical Philosophy* **18** (2009): 145-188.
[7] Hellman, G. *Mathematics without Numbers: Towards a Modal-Structural Interpretation* (Oxford University Press, 1989).
[8] Hellman, G. "Structuralism without Structures", *Philosophia Mathematica* **4** (1996): 100-123.
[9] Hellman, G. and Shapiro, S. "Towards a point-free account of the continuous", *Iyyun* (forthcoming).
[10] Lewis, D. *Parts of Classes* (Blackwell, 1991).
[11] Menger, K. "Topology without Points", *Rice Institute Pamphlets* **27** (1940): 80-107.
[12] Roeper, P. "Region-Based Topology", *Journal of Philosophical Logic* **26** (1997): 251-309.
[13] Roeper, P. "The Aristotelian Continuum. A Formal Characterization", *Notre Dame Journal of Formal Logic* **47** (2006): 211-231.
[14] Tarski, A. "Foundations of the Geometry of Solids", in *Logic, Semantics, and Metamathematics: Papers from 1923 to 1938* (Oxford, 1956).

Manuscript

Transfer and templates in scientific modeling

Abstract

The notion of (computational) template has recently been discussed in relation to cross-disciplinary transfer of modeling efforts and in relation to the representational content of models. We further develop and disambiguate the notion of template and find that, suitably developed, it is useful in distinguishing and analyzing different types of transfer, none of which supports a non-representationalist view of models. We illustrate our main findings with the modeling of technology substitution with Lotka-Volterra Competition equations.

1. Introduction. One intriguing feature of modeling techniques is that they may be applied across scientific disciplines. Harmonic-oscillator models, for instance, are seemingly applied wherever there is scientific work to be done. Still, not all models are migratory. The Nambu-Goldstone model, for instance, is a staple of quantum field theory, but sees no application elsewhere. The evaluation of modeling efforts in different contexts of application warrants further analysis – which minimally requires a clear identification of what may be *transferred* between such contexts. On the semantic view of models, for instance, transfer could concern the mathematical structure (i.e., a system of coupled differential equations), this structure along with its interpretation (i.e., the representation of a target system as a harmonic oscillator), or anything "in between".

Recently, (computational) *templates* have been proposed as the subjects of transfer (Humphreys 2002, 2004; Knuuttila 2009, 2011; Knuuttila and Loettgers 2012). Templates are types of differential equations, such as Lotka-Volterra equations, or modeling techniques, such as agent-based modeling, that are primarily constructed for their computational tractability, and that can be applied across disciplines to phenomena that, in the most extreme case, "may have nothing in common physically" (Humphreys 2002: S4). Remarkably, the notion of template has been used to argue for both a "selective" realist and a thoroughgoing instrumentalist view of modeling. In particular, the tractability-driven and "opportunistic" (Knuuttila 2009: 74) transfer of templates has been used to argue that models are epistemic tools, which are constructed and manipulated to contribute to a modeler's understanding, and not (primarily) valued for their representation of target systems. Thus, while it seems intuitive to claim that cross-disciplinary modeling

efforts[1] involve transfer of templates, there is a tension in (applications of) the notion of template with regard to one of the central philosophical questions regarding models.

In this paper, we argue that the notion of template illuminates cross-disciplinary modeling efforts, but that it does not support non-representationalism with regard to models. We distinguish three types of cross-disciplinary applications of modeling efforts, all of which may be understood as transferring templates and not interpreted computational models. Where templates are studied independently from computational models, there is only transfer in a degenerate sense; and where templates are genuinely transferred, this is strongly motivated by applications in computational models, valued in their different disciplinary contexts. Still, a marginal, but non-negligible role in modeling efforts for studying templates free from any specific interpretation shows that templates should not be identified with computational models. We illustrate our claims with a case study of transfer: the application of Lotka-Volterra Competition (LVC) equations in modeling technology substitution.

2. Computational and non-representational templates. The notion of computational template was proposed by Paul Humphreys (2002, 2004), in the context of emphasizing the importance to science, especially with regard to the interconnections between disciplines,[2] of computational techniques rather than

[1] Throughout, "cross-disciplinary modeling efforts" is used where we do not want to express commitment about any items (models or templates) that are transferred.

[2] We use "discipline" to indicate a "not necessarily large "branch of scientific research with characteristic subject matter and method(s) of inquiry.

theories. Mentioning several examples, including Laplace's and Lotka-Volterra equations and normal distributions, Humphreys argues that some modeling techniques see widespread use primarily because of their computational tractability. The notion of computational templates is meant to identify what is common to these applications. Humphreys distinguishes such templates from computational models: the latter come with an interpretation that relates a formalism to a specific subject or target system, whereas the former are relatively independent of any specific subject. Thus, "templates with different interpretations are not reinterpretations of the same model, but are different computational models entirely" (Humphreys 2002: S7). Thus, transfer of a modeling technique involves applying the template, not the model, to a new subject matter; there is, strictly speaking, no *model* transfer.

Humphreys warns against an instrumentalist conception of models, and claims that modelers take some parts of their models as true and others as false – in both cases expressing ontological commitment rather than the non-commitment that would indicate an instrumentalist attitude. Users of a template take a selective-realist attitude, by adding to the template (minimally) a subject-dependent *correction set*, which details the effects of relaxing its *construction assumptions* – the abstractions, idealizations, constraints and approximations that went into the construction of the template. Moreover, Humphreys maintains that construction of a template is not interpretation-free: the template is constructed in the light of its application to specific target systems, and at least one (subject-dependent) correction set is co-constructed.[3]

[3] "The correction set is also always subject-dependent and so, despite its flexibility, is the template itself. This is in part because of the inseparability of the template and its interpretation, in part because of the connection between the construction of the template and the correction set." (Humphreys 2002: S10).

Tarja Knuuttila uses the notion of template in her epistemic-tool account of models (e.g., Knuuttila 2009, 2011). On this account, models are primarily "result-oriented" instruments for increasing the modeler's understanding of the world. Models are purposefully constructed and manipulated, like tools; in particular, in evaluating a model, what matters is not its representational relation to target systems, but the contribution that its construction and manipulation makes to the realization of a given epistemic purpose (e.g., Knuuttila 2011: 263).

Knuuttila specifically mentions templates in her discussion of the "opportunistic" adoption of models constructed in other disciplines.[4] She emphasizes that the cross-disciplinary application of templates is guided first and foremost by considerations of tractability or solvability rather than any ability of the transferred item to represent accurately the (new) target system. She then uses the latter feature to argue against the view that models provide knowledge in virtue of being representational, intrinsically or as determined by modeler's intentions. The cross-disciplinary and opportunistic use of templates would show that modelers seek to "learn from the construction and manipulation of models quite apart from any determinate representational ties to specific real-world systems they might have" (Knuuttila 2011: 267). As "epistemic tools", models may provide understanding in a variety of contexts, none of which is prevalent over others in terms of intrinsic representational content or modeler's intention. Moreover, opportunism is recommended, not restricted, in the light of the result-orientedness of modeling: new

[4] "[T]here is an element of opportunism in modelling: the template that has proven successful in producing certain features of some phenomenon will be applied to other phenomena, often studied within a totally different discipline." (Knuuttila 2009: 74; 2011: 268).

applications of a template are to be evaluated on the basis of results obtained in the new context, not prior to transfer (Knuuttila 2011: 268). Here, a strong positive analogy with tools is employed: like tools, models and/or templates may be used for a variety of purposes, not all of which are foreseen by the tool's original designer, and many of which require fine-tuning or tinkering on the user's part before proving their true value to the purpose. The only explicit negative analogy is that models serve an epistemic, tools a practical purpose.[5]

Elsewhere, Knuuttila (with Loettgers, 2012) employs the notion of template to offer another, more implicit argument for her non-representationalism: templates can be constructed without any representational relation to a specific target system in mind. To establish this, the construction of Lotka-Volterra models by Volterra and Lotka is contrasted. The mathematical biologist Volterra was motivated by empirical phenomena regarding a specific target system (marine ecosystems), and only constructed a highly idealized set of coupled differential equations to model this phenomenon after concluding that a more realistic model would be mathematically intractable. By contrast, the general systems theorist Lotka derived the same set of differential equations from an abstract theory, irrespective of any specific target system, and only then showed that these equations could be applied to model oscillations in ecosystems and in concentrations of chemical substances. Thus, the very construction of a template may be non-representational – contrary to Humphreys' (2002) claim.

[5] This effectively restricts the analogy to a subclass of tools, since measuring instruments such as rulers and cognitive artifacts such as abaci do serve an epistemic purpose.

Summing up, the notion of template seems to offer a view of transfer of modeling efforts that is both plausible and at odds with representationalism or even realism regarding models. In what follows, we shall show that the template account of, in particular, transfer is in need of further development, which retains its plausibility and resolves the apparent tension with a representationalist view of models.

3. Cross-disciplinary modeling as transfer of templates. In the previous section, we noted that there is a tension in the notion of template: for one author, it supports a selective-realist, representationalist view regarding models, for another an instrumentalist non-representationalism. One might, in response, opt for abandoning the notion. To retain it, we shall in this section develop the notion in such a way that the tension is resolved.

To see why the notion needs developing, consider the claim that cross-disciplinary applications of modeling efforts involve templates, which are primarily valued for their computational tractability. As it stands this claim is uninformative. *That* something serves as a cross-disciplinary template does not make clear *why* some modeling efforts (e.g., involving coupled-oscillator models) see cross-disciplinary application and others (e.g., involving Nambu-Goldstone models) do not. Moreover, computational tractability cannot provide the sole reason: Nambu-Goldstone equations are as computationally tractable as Lotka-Volterra equations, but only the latter feature in cross-disciplinary templates.

Now, within each context of application of modeling efforts, computational tractability is valued because it allows derivation of specific implications or simulation of specific behaviors. Still, in transferring modeling efforts, both

disciplinary interests in and the interpretation of implications and behavior change. Balancing these evaluatively relevant aspects is difficult. On the one hand, emphasizing the versatility of the transferred items, and the necessary change of interpretation may underestimate the reason as to *why* tractability is still valued, viz., because of specific implications or behaviors, of new disciplinary interest – which may even be a direct counterpart of the original interests. Emphasizing similarities on the level of target systems, on the other hand, runs the risk of wrongfully equating templates and models and implying that, because representational content changes across applications, models are not primarily intended to represent target systems.

There is an ambiguity in the very notion of a template that is directly related to this balancing act. On one reading, it may refer to a purely formal object, the behavior of which can be studied independently of any context of application. On another reading, a template may be what computational models, valued in different disciplinary contexts, have in common. Although representational content is necessarily different in these contexts, this does not entail that the template is valued exclusively for its formal properties: the applicability of a template in one discipline may still be justified by reference to computational models in another discipline. Lotka's construction of the Lotka-Volterra template illustrates this ambiguity: it may be read as construction of a mathematical object, valued only for its tractability; or as a starting point for constructing multiple computational models, valued (also) for their diverse representational content.

To resolve the ambiguity and improve our understanding of the role of templates in modeling, we may distinguish three types of cross-disciplinary applications of modeling efforts. All of these may be understood as transferring templates, not computational models. Yet the motivations for these application, and

consequently the justification for use of the template, are relevantly different – and bring to light the role of interpretations in the evaluation of templates.

In the first type, modelers in one disciplinary context draw on modeling efforts in another discipline, not only because these involve application of computationally tractable mathematical structures (typically: sets of differential equations), but also – and primarily – because they want to apply the same implications of the computational models to similar target systems, or behavior of target systems. In such 'conformist transfer', not only the computational template is transferred, but also – in Humphreys' terms – its construction assumptions and correction set, appropriately re-interpreted, in order to transfer what is taken to be a central result. In justifications for this transfer, one would expect modelers to emphasize similarities between target systems, on the level of properties and/or behavior, at least as much as computational tractability. Possibly, but not necessarily, this emphasis on similarities takes the form of suggesting highly abstract models or encompassing theories, as is also acknowledged in cognitive theories of analogical reasoning (e.g., Holyoak and Thagard 1995).

In a second type of transfer, modelers draw on efforts in another discipline because they are interested in *different* implications of the same mathematical structure. For such 'creative transfer', a more general or extensive evaluation of the computational tractability of the template may be required, since the sensitivity of previously unstudied implications to construction assumptions must be assessed. This might lead to reformulation of the correction set. In justifications of creative transfer, one would not expect modelers to emphasize similarities between the properties and behavior of target systems and, by contrast, a stronger emphasis on formal analysis or general robustness of the template. However, this analysis is not independent from an

interpretation of the template in its new context of application: its ability to represent behavior of a target system motivates application of the template, even if this behavior may have no counterpart, or no counterpart of any disciplinary interest, in the original context of application. In the extreme, target systems may have "nothing in common physically".[6] Here, what is transferred is a template plus interpretation potential.

Extremely creative transfer must be distinguished from a third type of extension of modeling efforts, which can only be called "transfer" in a degenerate sense. Here, modelers may study the behavior of a mathematical structure that has seen application in one disciplinary context *purely* out of an interest in its computational tractability or its formal implications. They may, for instance, relax various constructive assumptions or change parameter settings, not in order to make a computational model more realistic, but because they want to test the general robustness of a template. Here, the template is studied independently of any context of application, even the original one – these investigations strictly speaking study *template* behavior, not *model* behavior. Moreover, although these modeling efforts may not involve transfer, they may prove valuable in justifying subsequent creative transfer, and may (but need not) be motivated by the possibility of such transfer.

Templates are thus involved in a variety of modeling efforts, and only seldom independently from (the presentational content of) computational models, valued in their different disciplinary contexts. Where templates are studied in independence from computational models, there is only transfer in a degenerate sense. Thus, although templates are strictly speaking without representational content, and they are

[6] Note just how extreme such a case is, since similarities must be absent (or remain unmentioned) on the level of entities, properties, relations and behavior. Analogical reasoning must, in short, play no role whatsoever in such transfer of modeling efforts.

what is transferred, the phenomenon of transfer can hardly be used in support of a non-representationalism regarding models. Still, that there is a role in modeling efforts for studying template behavior free from any specific context of application shows that templates should not be identified with computational models.

4. A case study: LVC models of technology substitution. In this section, we look at one case of transfer of modeling efforts: the modeling of processes of technology diffusion with Lotka-Volterra Competition (LVC) equations. We first give a necessarily brief description of these modeling efforts in their disciplinary context. Then, we analyze some features of these efforts with the notion of template, as it was developed and differentiated in the previous section. In particular, we point out a distinction between conformist and creative transfer, the importance of application of templates in computational models, and a marginal (but non-negligible) role of interpretation-free templates.

Predicting and explaining how technological innovations capture market share is of obvious commercial interest. One model of this process fits the simplest logistic curve to the growth rates of technologies (Fisher and Pry, 1971), following the observation that these rates tend to follow a sigmoid curve after capture of a small but significant market share. The (perhaps surprising) predictive success of these and other phenomenological models has led to widespread use in industry, and to an increasing focus in research on hybridization of existing models for predictive purposes,[7] as well as attempts to construct more explanatory models.

[7] Meade and Islam (1998) review twenty-nine phenomenological models and show that a combination provides a better fit to data sets than each of the individual models.

One such attempt follows a suggestion by Fisher and Pry that the diffusion of innovations can be understood as, primarily, a process of competition between an emerging and an established technology. Several researchers have therefore, for explicitly explanatory purposes, sought to apply the LVC equations to the growth rates of rival technologies. They describe the merits of these models as providing "clearly defined assumptions about the nature of technological growth" (Porter et al. 1991: 197). Often, the behavior of the LVC model is studied in relation to the various phenomenological models, for instance by arguing that, under a range of conditions, LVC models reduce to Fisher-Pry models (Bhargava 1989). Occasionally, LVC equations are directly fitted to data sets of competing technologies. Farrell (1993), for instance, applies them to the substitution of soldered cans by lead-free cans, of fountains pens by ballpoint pens; and two other substitution processes.

These modeling efforts are thus explicitly motivated by explanatory concerns, expressed in claims that LVC models should provide an understanding of the mechanisms of technology substitution. Moreover, the analytic and computational tractability of these models plays an important role, in deriving well-established phenomenological models as special cases (e.g., Bhargava 1989), in deriving general properties of systems of competing technologies (e.g., Saviotti and Mani 1995), or in applications to data sets (e.g., Farrell 1993).

Yet there are at least two strategies for seeking this understanding, reflecting the distinction between conformist and creative transfer made in the previous section.

The first strategy – explicit in, e.g., Bhargava 1989; Porter et al. 1991; Farrell 1993 – starts from noting the similarity between the logistic (Pearl-Verhulst) growth models of ecology and the Fisher-Pry model, where a "technological" counterpart is indicated for each element of the biological model: technologies are likened to yeast

cultures, growing in an environment with a maximum carrying capacity (corresponding to market saturation), etc. Then, it is noted that LVC models should comprise Fisher-Pry models as a special case, just as they comprise logistic growth models, the technological interpretation of the latter is carried over to the LVC equations, and the behavior of these equations that is familiar from biological applications (e.g., a bell-shaped growth curve for the □defending□species/technology on emergence of a new species/technology) is found in data sets on technology substitution.

A second strategy □explicit in Saviotti and Mani (1995) □involves the same template, but strays further from its ecological context of application. It starts by constructing a model that is supposed to capture the microeconomic mechanisms behind technology substitution: a set of three equations with an elaborate, detailed interpretation in terms of obsolescence, learning-by-doing, and purchase of intellectual property rights and other factors that have no obvious counterpart in ecology □and even for those factors that do, no such counterpart is mentioned. The behavior of these equations is not studied, apart from a qualitative reconstruction of various modes of competition (perfect, monopolistic, Schumpeterian and inter- and intra-technological), mostly known from the economic literature. Only then, the LVC equations are introduced, as an □aggregate representation□of technological change, with reference to their similar status for ecological change. After some manipulations, counterparts of the microeconomic model □especially of the parameters corresponding to its distinction between inter- and intra-technological competition □ are sought; and the behavior of the manipulated equations is simulated to derive a relation between technological variety and the relative strength of modes of competition, along with the conditions under which the relation holds.

Both strategies involve transfer of the same template, and in each case, its adoption is partly motivated by its tractability (analytical or computational) and partly by its interpretability in technological terms. However, the first strategy may be identified as strongly conformist, and the second as comparatively creative. This is revealed both in the interpretation of the template and in what is presented as its relevant behavior and assumptions. The first strategy attempts a term-for-term translation, and emphasizes behavior that is familiar from applications in ecology.[8] The second interprets the LVC equations in the same terms as a microeconomic model, and studies behavior that has no obvious counterpart in ecological applications.[9] This difference also shows in remarks made on the sensitivity of results: those that follow the first strategy note that applications of the LVC equations assume a stable environment, and find a counterpart in fixed-sized markets; the second strategy involves explicit analyses of conditions under which the main results obtain, formulated as relevant parameter intervals and ceteris paribus conditions. This confirms, with qualifications, Humphreys claim that conditions on the applicability of the template equations do not feature as ceteris paribus conditions in statements of results: they do not feature as such in conformist transfer, but they do in creative use.

Another feature of templates that is revealed in LVC modeling of technology substitution is that transfer of the LVC template is strongly motivated by its application in (fully interpreted) computational models. In the first strategy,

[8] Farrell (1993) also seeks to translate the *method* of applying LVC equations, in order to arrive at familiar results.

[9] Saviotti and Mani (1995) do note in passing that one of their intermediate results has an ecological counterpart.

technologically meaningful counterparts of virtually all ecological concepts are identified before presentation of the result – which is itself a counterpart of a central result in ecology. Thus, there is no discernable study of the behavior of the equations apart from a prior, and heavily 'bio-inspired' interpretation. The second strategy differs from the first, not in being interpretation-free, but in interpreting all concepts, as well as the central result, in micro-economic terms. Still, a tension between interpretability and tractability shows up occasionally. Most notably, Farrell (1993: 174) cautions against interpreting the interaction terms in the LVC equations in terms of comparative technological performance. Such an interpretation, while tempting, would neglect that '[t]here is no specific mechanism behind these *equations*' (emphasis added). Here, the formal character of the template is emphasized in order to prevent over-interpretation of the equations.[10]

Finally, in only one place, we find evidence for some interpretation-free manipulation of the LVC template in the literature on technology substitution. Morris and Pratt (2003) use a rather sophisticated graphical method to derive analytically that the LVC equations may 'revert' to the Fisher-Pry curves, but that they can only mimic, not match, the behavior of other phenomenological models. Although the positive result is the same as in papers that exemplify the first strategy, it is here derived without any interpretation of either the LVC or the Fisher-Pry equations – and the same goes for the negative result, which is unique to this paper.

[10] Farrell goes on to speculate about the possibility to derive a technological model from knowledge of the underlying mechanisms – which seems exactly what Saviotti and Mani (1995) claim to have done, arriving again at the LVC equations, which are now fully (micro-economically) interpreted.

There is, summing up, hardly any evidence for an interpretation-free application of templates, let alone for non-representational models; in neither of the two strategies for transferring the LVC template, the template is applied in isolation from computational models. Moreover, the representational content of these models sometimes including a translation of this content from other contexts of application is emphasized by practitioners in their attempts to understand the mechanism(s) of technology substitution. Still, we identified a marginal, but non-negligible role in these modeling efforts for studying the LVC template free from any specific interpretation, illustrating that templates should not be identified with highly abstract computational models.

REFERENCES

Bhargava, S.C. 1989. Generalized Lotka-Volterra Equations and the Mechanism of Technological Substitution. *Technological Forecasting and Social Change* 35:319-326.

Farrell, Christopher J. 1993. A Theory of Technological Progress. *Technological Forecasting and Social Change* 44:161-178.

Fisher, J.C. and R.H. Pry. 1971. A Simple Substitution Model of Technological Change. *Technological Forecasting and Social Change* 3:75-88.

Holyoak, Keith J. and Paul Thagard. 1995. *Mental Leaps*. Cambridge, MA: MIT Press.

Humphreys, Paul. 2002. Computational Models. *Philosophy of Science* 69:S1-11.

 2004. *Extending Ourselves*. New York: Oxford University Press.

Knuuttila, Tarja. 2009. "Isolating Representations vs. Credible Constructions?" *Erkenntnis* 70:59-80.

___ 2011. "Modelling and Representing." *Studies in History and Philosophy of Science* 42:262-271.

___ and Andrea Loettgers. 2012. "The productive tension." Forthcoming in: Paul Humphreys and Cyrille Imbert, eds., *Models, Simulations and Representations*. London: Routledge.

Meade, N. and T. Islam. 1998. "Technological Forecasting – Model Selection, Model Stability and Combining Models." *Management Science* 44:1115-1130.

Morris, Steven A. and David Pratt. 2003. "Analysis of the Lotka-Volterra Competition Equations as a Technological Substitution Model." *Technological Forecasting and Social Change* 70:103-133.

Porter, Alan L., A. Thomas Roper, Thomas W. Mason, Frederick A. Rossini and Jerry Banks. 1991. *Forecasting and Management of Technology*. Hoboken, NJ: John Wiley.

Saviotti, P.P. and G.S. Mani. 1995. "Competition, Variety and Technological Evolution." *Journal of Evolutionary Economics* 5:369-392.

Philosophy of Science

Relevance, not Invariance, Explanatoriness, not Manipulability: Discussion of Woodward on Explanatory Relevance.

--Manuscript Draft--

Manuscript Number:	
Full Title:	Relevance, not Invariance, Explanatoriness, not Manipulability: Discussion of Woodward on Explanatory Relevance.
Article Type:	PSA 2012 Contributed Paper
Keywords:	explanation; explanatory relevance; explanatory depth; invariance; Woodward; causal model of explanation; control; manipulability
Corresponding Author:	Cyrille Thomas Imbert Archives Poincaré, CNRS, Université de Lorraine Nancy, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Archives Poincaré, CNRS, Université de Lorraine
Corresponding Author's Secondary Institution:	
First Author:	Cyrille Thomas Imbert
First Author Secondary Information:	
Order of Authors:	Cyrille Thomas Imbert
Order of Authors Secondary Information:	
Abstract:	In Woodward's causal model of explanation, explanatory information is information that is relevant to manipulation and control and that affords to change the value of some target explanandum variable by intervening on some other. Accordingly, the depth of an explanation is evaluated through the size of the domain of invariance of the generalization involved. In this paper, I argue that Woodward's treatment of explanatory relevance in terms of invariant causal relations is still wanting and suggest to evaluate the depth of an explanation through the size of the domain of circumstances that it designates as leaving the explanandum unchanged.

Manuscript

Relevance, not Invariance, Explanatoriness, not Manipulability:

Discussion of Woodward on Explanatory Relevance.

Word count :

- abstract: 99 words

- article: 4924 words + 1 figure

Additional spaces have been included in the equations to make them more readable, which increased the word count by roughly 100 words.

Abstract

In Woodward's causal model of explanation, explanatory information is information that is relevant to manipulation and control and that affords to change the value of some target *explanandum* variable by intervening on some other. Accordingly, the depth of an explanation is evaluated through the size of the domain of invariance of the generalization involved.

In this paper, I argue that Woodward's treatment of explanatory relevance in terms of invariant causal relations is still wanting and suggest to evaluate the depth of an explanation through the size of the domain of circumstances that it designates as leaving the *explanandum* unchanged.

Relevance, not Invariance, Explanatoriness, not Manipulability:

Discussion of Woodward on Explanatory Relevance.

1. Introduction

The question of explanatory relevance has been for long a challenge for theorists of explanation. It is well-known for example that Hempel's DN model, Salmon's SR model or Salmon's causal models fail to characterize philosophically what type of information is relevant to the explanation of some fact *F* and should therefore figure in its explanation.

In the last two decades, James Woodward has developed a manipulationist model of explanation, which seems to fare better than its predecessors about explanatory relevance, if not to solve the issue, and that accounts for many of the usual tricky cases. In this model, explanatory information is information that is relevant to manipulation or control and that affords to change the value of some target *explanandum* variable by intervening on some other. Accordingly, the depth of an explanation is evaluated through the size of the domain of invariance of the generalization involved.

In this paper, I argue that Woodward's treatment of relevance in terms of invariant causal relations is still subtly but unavoidably wanting because it forces one to include within the explanation of a fact F much information that may be relevant to account for other facts of a same physical type but may be irrelevant to F. I further suggest to evaluate the depth of an explanation through the size of the domain of circumstances it describes as leaving the *explanandum* phenomenon unchanged.

In section 2, I briefly present Woodward's account of explanation and his notion of explanatory depth. I develop at length in section 3 a test case example dealing with the explanation of the law of Areas and describe two ways to explain this physical regularity. I show in section 4 that, whereas the first explanation includes clearly irrelevant facts, according to Woodward's account, it cannot be said to be less explanatory than the second. I further analyze why satisfying the manipulability requirement may imply to include irrelevant facts in explanations in order to make them deeper (in Woodward's sense). I further describe in section 5 a new criterion for judging explanatory depth and argue that this criterion and Woodward's criterion are incompatible. I finally emphasize in section 6 that manipulability is still a virtue, even if not an essential virtue of explanations and that, depending on the circumstances, one may be interested in developing explanations that are less explanatory (because they contain irrelevant facts) but that afford to control physical systems.

2. Woodward's manipulationist account of explanation

It may seem weird to challenge Woodward (and Hitchcock) on the question of explanatory relevance for they have themselves showed much acumen in diagnosing where existing accounts fail and offered new answers to the problem. Indeed, in his 1995 article, Hitchcock elegantly shows that the problem of explanatory relevance is still a worry for Salmon's causal model because identifying all the intermingled spatio-temporal causal processes running in some physical circumstances falls short of indicating why exactly some phenomenon takes place in these circumstances. As Woodward further notes, even if the right causal processes are identified, "features of a process *P* in virtue of which it qualifies as a causal process (ability to transmit mark *M*) may not be the features of *P* that are causally or explanatorily relevant to the outcome *E* that we want to explain" (Woodward, 2003, 353).

In this context, it comes as no surprise that Woodward tries to answer the above worries by

means of his causal model. Doing justice to all aspects of Woodward's rich treatment of explanatory relevance and explanation would take much longer than can be done within this short paper. The next paragraphs are therefore merely devoted to reminding the reader some important aspects of Woodward's account so that what it amounts to when it comes to the analysis of the coming example appears clearly.

For Woodward, "explanation is a matter of exhibiting systematic patterns of counterfactual dependence" (2003, 191). Explanatory generalization used in an explanation must indicate that the *explanandum* was to be expected *and* how it would change, were some changes made in the circumstances that obtained; said differently, good explanations "are such that they can be used to answer a range of counterfactual questions about the conditions under which their *explananda* would have been different" (ibidem).

In this perspective, "explanatory relevant information is information that is potentially relevant to manipulation and control" (2003, 10). In other words, something is relevant information if it essentially figures in an explanation describing how the *explanandum* was to happen and how it would change, were the properties described in the *explanans* modified. This requirement also discards irrelevant circumstances through the identification of irrelevant variables: "an *explanans* variable *S* is explanatorily irrelevant to the value of an *explanandum* variable *M* if *M* would have this value for any value of *S* produced by an intervention" (2003, 200).

Woodward further defines the notion of invariance of a generalization. A generalization can be stable under many changes of conditions not mentioned in it. For example, Coulomb's law holds under changes in the weather. By contrast, a generalization that "continues to hold or is stable in this way under some class of interventions that change the conditions described in its

antecedent and that tells us how the conditions described in its consequent would change in response to these interventions is invariant under such interventions" (1997, S.31)[1].

It is then clear that invariance is a gradual notion because a generalization can hold under more or less interventions. Accordingly, depending on the degree of invariance of the generalization they rely upon, explanations provide patterns for answering more or less what-if explanatory requests about these counterfactual circumstances and therefore for controlling the corresponding systems.

Woodward further claims that the concept of invariance provides a means for evaluating the goodness of explanations – what he calls "explanatory depth": "We can thus make comparative judgments about the size of domains of invariance and this is all that is required to motivate comparative judgments of explanatory depth of the sort we have been making" (1997, S.39). To put things briefly, the more invariant, the more explanatory, or to use Woodward's own words: "generalizations that are invariant under a larger and more important set of changes often can be used to provide better explanations and are valued in science for just this reason" (2003, 257).

At this step, my claim can be precisely formulated: even if they are valued in science, more invariant explanations are not always more explanatory because the request for invariance may run contrary to the fundamental request for relevance that explanations should primarily satisfy.

3. The law of Areas and its explanations

[1] More precisely, invariance is defined by means of the notion of "testing intervention". See (2003, 250) for more details.

The test case I now want to investigate is the explanation of the law of Areas (also called "Kepler's second law"), which states that, "for planets in our solar system, a line joining a planet and the sun sweeps out equal areas during equal intervals of time". I shall describe two explanations of it and compare them with respects to invariance and relevance.

As we shall see, the first explanation (hereafter explanation 1) relies upon the general angular momentum theorem. Let us go deeper into it. Let us assume a Galilean reference frame, a fixed axis M' with position given by vector **r'** and a moving material point with position given by vector **r**, having mass *m* and momentum **p** (bold characters denote vectors). The angular momentum of M about M' is defined by: $\mathbf{L}_{r'} = \mathbf{(r'-r)} \times \mathbf{p} = m\,(\mathbf{r'-r}) \times \mathbf{v}$, where the symbol "×" stands for the usual external product. Let **F** denotes the sum of forces applied to M. The momentum of **F** about axis M' or torque is defined as $\boldsymbol{\mu}_{\mathbf{F/M'}} = \mathbf{(r'-r)} \times \mathbf{F}$. Then, deriving the angular momentum yields

$$\frac{d\mathbf{L}_{r'}}{dt} = \frac{d((\mathbf{r}-\mathbf{r}')\times\mathbf{p})}{dt} = (\mathbf{r}-\mathbf{r}')\times\frac{d\mathbf{p}}{dt} + \frac{d(\mathbf{r}-\mathbf{r}')}{dt}\times\mathbf{p}$$

Because the momentum **p** is collinear to the speed of M, the second term in the right-hand part of the equation is null. So far no physics has been used. Newton's second law says that $d\mathbf{p}/dt = d(m\mathbf{v})/dt = m\mathbf{a} = \mathbf{F}$. So finally, one gets

$$(1)\quad \frac{d\mathbf{L}_{r'}}{dt} = (\mathbf{r}-\mathbf{r}')\times\mathbf{F} = \mu_{\mathbf{F/M'}}$$

For a collection of particles, one can also define the total torque $\mu=\Sigma\,\mu_i$, which is the sum of the torques on each particle, as well as the total angular momentum **L**, which is the sum of momentum of each particle and one gets

$$(1.5)\quad \mu=\Sigma\,\mu_i = d\mathbf{L}/dt.$$

The total torque is the sum of the momentum of all forces, internal and external. But, because of Newton's law of action and reaction, the torques on two reacting objects compensate and therefore, the internal torques balance out pair by pair. In conclusion, "the rate of change of

the total angular momentum about any axis is equal to the external torque about that axis". This is the general angular momentum theorem, which is true for any collection of objects, whether they form a rigid body or not.

If one wants to explain the law of Areas, one should finally note that, in the case of the Earth/Sun two-body system, if $\mathbf{v_E}$ denotes the speed of the Earth, $\mathbf{r_E}$ its position, $\mathbf{F_G}$ the gravitational force, $\mathbf{L_E}$ the Earth momentum about the Sun, α the angle between $\mathbf{r_E}$ and $\mathbf{v_E}$, and A_E (t) the swept area in function of time, in virtue of the definition of the outer product,

$$(2)\ \frac{\|\mathbf{L_E}\|}{m_E} = \frac{\|\mathbf{r_E} \times \mathbf{v_E}\|}{m_E} = \|\mathbf{r}_E\|.\|\mathbf{v}_E\|\sin(\alpha) = 2.\frac{dA_E(t)}{dt}.$$

Because this relation holds for each mass point, the relation $\mu=\Sigma\ \mu_i = d\mathbf{L}/dt$ can now be seen as describing the variation of the variation of the sum of the areas swept by each point of a system about an axis, be it a rigid body or a set of independent mass points.

In the case of the Earth-Sun system, it should further be noted that the momentum of the gravitational force $\mathbf{F_G}$ about the Sun is zero (because the force and the vector **r** are collinear). Therefore, because of (1.5), the angular momentum of the Earth about the Sun is constant and because of (2), A(t) grows linearly with time, which demonstrates that the law of Areas obtains.

This explanation perfectly fits Woodward's account of explanation and one can repeat what he says about his paradigmatic case of the theoretical explanation in terms of Coulomb's law of the electrostatic relation $E=\lambda/(2\pi\varepsilon_0 r)$ (203,196-204). The explanation does exhibit the features emphasized by DN theorists: it is a deductively valid argument in terms of Newton's second law and the description of the system (positions, speeds and masses of the points, forces). But in addition, it does exhibit a systematic pattern of counterfactual dependence, which can be summarized by combining (1.5) and (2) into the general relation (3) $\mu=\Sigma\ \mu_i = dL/dt=2\ \Sigma\ m_i\ d/dt\ (dA_i(t)/dt)$, which the law of Areas is a special case when the right variables

are assigned the right values (two bodies, one central gravitational force, etc.). The derivation describes how the *explanandum* law of Areas would change according to (3) and how it systematically depends on Newton's second law, the forces and the particular conditions cited in the *explanans*. More specifically, the explanation makes clear how the total swept area would vary were the mass, speed, position of the Earth different, were additional forces at play but also were additional bodies included in the system. In short, (3) and the explanation including it also indicate how to answer a range of what-if questions about counterfactual circumstances in which the *explanandum* would have changed. Regarding the range of these questions and the invariance of the explanation, it is difficult to do better, because Newton's law and (3) cover all situations in classical physics and therefore all classical changes that can be brought about to the two-body system case.

Let us now turn to the second explanation (hereafter explanation 2). In order to give the reader a clearer feeling of why it is better, I shall give two versions of it, one of which more pictorial. Let us start with the vectorial derivation. Because of relation (2), the law of Areas obtains if the intensity $\|\mathbf{L_E}\|$ of the angular momentum $\mathbf{L_E}$ of the Earth about the Sun is constant. In virtue of relation (1), this happens when $\mathbf{(r'-r)} \times d\mathbf{p}/dt = 0$, which is the case if $d\mathbf{p}/dt$ and $\mathbf{(r'-r)}$ are collinear. This is so because the only force at play is radial and the variation of momentum of a particle is along the direction of the force exerted upon it, that is $d\mathbf{p}/dt = \alpha\ \mathbf{F}$, where α is real, not necessarily constant and not specified. Newton provides a more geometrical way to see the explanation:

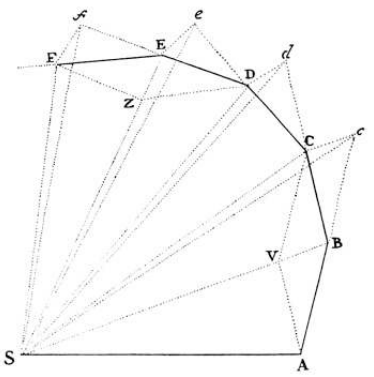


Figure 1: Geometrical demonstration of the Law of Areas by Newton (1726/1972)

The Earth's trajectory goes through A, B, C, etc. and the law of Areas obtains if the area of SAB, SBC, etc. are numerically equal. The explanation of each trajectory step is decomposed in two parts. On the one hand, if no force was at play, in virtue of the inertia principle, the Earth would go straight from B to c in one time interval with AB=Bc. This implies that the area of SAB and SBc are numerically equal. On the other hand, if the Earth was motionless in B, because of the central gravitational force, it would go somewhere on (SB), say in V. By combining the two moves, the Earth finally goes to C, with **BV**=**cC**. Because (Cc) and (SB) are parallel, the area of SAC and SBc are also numerically equal. By combining the two equalities, one gets that that the area of SAB and SAC are numerically equal. The law of Areas finally obtains by taking smaller and smaller time intervals. The important point is that the numerical equality between the area of SBc and SBC obtains whatever the position of V on (SB): in other words, it obtains provided that the change of momentum due to a force is along the force direction, that is, provided $d\mathbf{p}/dt = \alpha\ \mathbf{F}$.

How good is this second explanation? First, it also exhibits the features emphasized by DN theorists: it is a deductively valid argument in which some nomological component is essentially needed (as well as the description of some particular circumstances). It shows in

addition that the whole content of Newton's second law is not required within the explanation. More precisely, the quantitative part of Newton's second law, which relates the values of forces and acceleration, can be removed for the premises without altering the validity of the argument. Better, from a physical point of view, this removal brings some important piece of explanatory information because it indicates more specifically what in the physics is essential for the law of Areas to obtain. The quantitative aspect of the momentum variation is shown to be explanatorily irrelevant, which indicates that the law of Areas does obtain for all worlds with a dynamical law such that the variation of momentum is along the force direction – and this is a piece of explanatory information that explanation 1 does not provide because it includes the described irrelevant information.

Accordingly, explanation 2 is also instrumental to answer what-if questions about what would happen should the intensity of the force be different, time be discrete or the gravitational constant change with time. So, the corresponding explanatory generalization is also invariant under a large range of interventions.

4. Comparison between the two explanations regarding depth and diagnosis about the inadequacy of Woodward's account

Let us now see how the two explanations comparatively fare according to Woodward's criterion of explanatory depth. As just mentioned, both explanations are invariant under a large range of interventions. As we saw, Woodward suggests assessing explanatory depth by comparing domains of invariance. In the present case, none of the two explanations can then be said to be deeper than the other because none of the two sets is a subset of the other. Indeed, explanation 1 directly yields answers to what-if questions about how the total swept area quantitatively changes when, say, non radial forces are at play or more bodies involved, which explanation 2 does not (because it omits the quantitative part of Newton's second law).

Conversely, explanation 2 explicitly indicates that the law of Areas would still obtain in circumstances in which Newton's second law would be violated, which explanation 1 does not, because it designs as explanatory relevant the whole law with its quantitative aspect. Overall, from Woodward's perspective, we have a situation with two good explanations which explanatory depth cannot be compared because their domains of invariance only partly overlap. And this is a case that is accommodated by Woodward when he notes that the comparison of the domains of invariance of explanations "obviously yields only a partial ordering" because "for many pairs of generalizations, neither will have a range of invariance that is a proper subset of the other" (2003, 262-64).

My point is that this woodwardian conclusion is not satisfactory: if one focuses upon the relevance of the explanatory material regarding the *explanandum*, explanation 2 is better than explanatory 1. It is indeed commonly agreed that an explanation of A should merely include explanatory information that is relevant to the occurrence of A (at least if one's epistemic goal is to provide an explanation of A that is as explanatory as possible (see section 6 for more comments about this restriction). As mentioned earlier, explanation 2 omits explanatory material that is irrelevant to the occurrence of the law of Areas, which explanation 1 does not. It is then no surprise that explanation 1 provides an answer to many what-if questions which answer depends on this irrelevant material and cannot therefore be given by explanation 1. However, while these additional answerable questions contribute to extend the invariance of explanation 1, the ability to answer them should not be seen as a sign of the greater goodness of explanation 1 (quite the contrary!) because, as the Newtonian investigation described

above shows, answering them requires some causal information that is here explanatorily irrelevant[2].

Let us now try to see more clearly why Woodward's account leads to include irrelevant features in explanations to make them deeper. The reason seems to be that he requires that an explanation should account for many counterfactual cases that belong to a same physical type, defined in terms of the *explanandum* variable appearing in the explanatory generalization, and which the *explanandum* fact is an instantiation of. But this compels him to include in the explanatory material not only the facts that are explanatorily relevant to the target *explanandum* but also the facts that are explanatory relevant to all the values the *explanandum* variable may take. But as the example shows, the explanatorily relevant facts for the latter and the former need not coincide. The moral to draw is that facts belonging to an identical type do not always have the same explanations nor explanations of the same type.

Here, it is important to note that the *explanandum* type that requires to draw this moral (the variation of the swept area) is not the product of some gerrymandering artificially associating pears and apples. So the moral should be rephrased more precisely and strongly like this: facts belonging to an identical *bona fide physical* type (corresponding to the *explanandum* variable of a genuine physical generality) do not always have the same set of explanatory relevant facts nor explanations of the same type.

This conclusion has a counterpart in terms of whether domains of invariance are appropriate to assess the depth of an explanation and which what-if erotetic requests are *appropriate* for this task (to use a notion Woodward often relies upon). Requiring that an explanation of a

---

[2] Of course, these irrelevant features belong to a fundamental causal law, which is true in all models described by classical physics. But this does not imply that they should pop up in all our explanations of physical phenomena.

target *explanandum* fact F should allow one to answer what-if questions about counterfactual circumstances corresponding to the invariance domain of some general and functionally described regularity, which the *explanandum* case is an instance of, may imply to include in the explanatory material physical information that is relevant for these circumstances but not for F. Accordingly, even if these explanatory requests are by themselves scientifically legitimate, it may be illegitimate to judge the goodness or depth of an explanation of F by the ability it provides to answer these requests because the physical information necessary for this task may be explanatory irrelevant regarding *F* - and this information should therefore not be included in a good explanation E of F, which removes the possibility of answering these requests on the basis of E. In short, being a what-if question about some circumstances in the domain of invariance of the explanatory generalization that one uses in the explanation E is not a sufficient condition for being an appropriate question for testing the depth of E because this criterion is incompatible with a satisfactory treatment of the problem of relevance for explanations.

The conclusion regarding the evaluation of explanatory depth in terms of domain of invariance comes naturally. It is not legitimate to evaluate the depth of an explanation by assessing the domain of invariance of the generalization used in it. Performing well on the invariance criterion leads to promote explanations of individual facts that are special cases of general explanatory patterns built on generalizations that are invariant on large domains… but it potentially also leads to violate the requirement of relevance for the explanations of these individual facts.

5. Another criterion for explanatory depth

Still, as can be inferred from the discussion of the example, it seems that a good explanation (which satisfies the criterion of explanatory relevance) does provide answers to many

appropriate what-if questions. Explanation 2 shows that the law of Areas would still obtain in many circumstances in which the quantitative part of Newton's second law or the intensity of the gravitational law would be different. It thereby enables one to answer in the affirmative the corresponding "would-the-explanandum-still-be-the-case" (in short "would-still" questions). For a derivative explanation, this set of circumstances in which the *explanandum* is shown by an explanatory argument to be left unchanged corresponds to the set of situations in which the premises of the explanatory argument are true. Further, the more irrelevant information is removed from the premises, the weaker these explanatory premises and the wider the class of situations to which they apply. Let us call this class of situations the domain of strict invariance of the explanation (by contrasts with Woodward's notion of domain of (large) invariance of the generalization employed in the explanation, hereafter "large invariance"). Then, the above discussion leads to the following suggestion:

(S) The wider the domain of strict invariance of an explanation, the deeper the explanation.

It would take much more that can be said here to develop this suggestion into a fully-fledged proposal about the nature of explanation. In particular, a critical comparison with notions discussed by Reichenbach or Salmon in different contexts such as the notions of *broadest homogeneous reference class*, *maximal class of maximal specificity* or *exhaustiveness* (Salmon, 1989, 69, 104, 193) would be helpful. Nevertheless, the following remarks are in order. First, (S) indicates how an explanation can be turned into a better one by expurgating its premises from irrelevant information; but it does not however indicates in general what type of information can be present in the premises for something to count as a potential explanation. Therefore, it should not be seen as something standing on its own (otherwise, the best explanation would be the self-explanation of one fact by itself). Second, the domain that is here described should be distinguished from the scope of the laws or the

domain of invariance of the generalization present in the premises, which characterize statements: strict invariance characterizes the explanation itself. Alternatively it can be seen as the domain of the explanatory generalization saying that when the premises hold (in this or different worlds), so does the *explanandum*. Third, just as for Woodward's account, this criterion is likely to describe only a partial order over explanations. Finally, it should be noted that the criteria of having a large domain of large invariance and of having a large domain of strict invariance go into two opposite directions. Indeed, explanations with large domains of general invariance require generalization with much physical information packed in it; whereas explanations with large domains of strict invariance require premises with as little physical information as possible in their premises. So it does not seem possible to try to conciliate both criteria about the nature of explanatory depth.

6. Concluding remarks: generality and manipulability *versus* specificity and relevance or the contextual choice of epistemic virtues in scientific practice

I have criticized in this article the use of the size of the domains of invariance of the generalizations used in explanation to describe the depth of these explanations. I have argued that this characterization of the goodness of explanations fares badly by the requirement of relevance, which explanatory explanations should primarily satisfy. To describe the goodness of explanations I have proposed a different criterion based on the notion of strict invariance and the ability to answer "would-still" questions offered by explanations. And I have emphasized that satisfying one criterion may run contra the satisfaction of the other.

One final word of caution is needed here. The above analysis dealt with the explanatory character of explanations of specific individual facts, which relevance is a clear component of. Now, like all other things, explanations may also have unspecific additional virtues, which may be philosophically unessential to them but practically crucial to their use. In the present

case, having a wide large invariance is no doubt such an unessential virtue. Indeed, an explanation with wide large invariance, even if it is of average quality regarding explanatory relevance, does provide a functional pattern for a family of similar explanations: it offers the opportunity to explain many similar phenomena with the same pattern of reasoning, which yields some significant economy of scientific and cognitive means. As any versatile tool, because it is general, such an explanation may prove useful, even if it is not optimal for specific explanatory tasks. Finding such explanations is therefore a scientifically legitimate (and difficult) task.

So should scientists favor in practice specific relevant explanations with wide domains of strict invariance over general explanations with wide domains of large invariance? I think there is no general answer to this question. *Pace* the philosophical interest for essential epistemic virtues, contextual interests are to prevail depending on what scientific needs are. Suppose that you are interested in controlling optical rays within optical fibers or the trajectory of a car in various circumstances; then there is little doubt that you will be interested in finding explanations with wide domains of large invariance so that you can determine how the rays or the cars will behave in a wide range of circumstances with one single functional relation and control them by adopting the external forcing. For some of these covered circumstances, it is likely that this single functional relation will contain unnecessary (irrelevant) information and for some specific cases you may even be using a sledgehammer to crack a nut; but why should you care? For control purposes, it may be more convenient to use one single relation covering all cases than a cumbersome wealth of them, each specifically targeted at some subset of circumstances.

Suppose now that you are interested is observing a green flash effect (some optical phenomena occurring after sunset or before sunrise, when a green spot is visible above the sun). Then, what you want to learn about the circumstances in which you stand a good choice

to observe a green flash effect and you want to know a set of circumstances that is as large as possible. Therefore, knowing which circumstances will not alter the phenomenon (because they are irrelevant to the mechanism involved) is crucial. In this case, you will be interested in discarding from the explanation any irrelevant information that restricts your knowledge of this set, even if it comes at the price of leaving out of the *explanans* physical information that may be useful to answer questions about what would happen in close circumstances (in which no green flash effect is observed). So you may end up with an explanation that is not useful for manipulationist purposes because it is specifically targeted at the green flash effect; perhaps this explanation will not even have a functional form (like above the explanation 2 of the law of Areas); but, because its *explanans* only describes the physical facts that are crucial for the green flash effect to happen and discards the other, it will be more explanatory and therefore more informative about the whole range of circumstances in which the observation can be made.

In conclusion, Woodward's criterion for explanatory depth seems more appropriate to characterize explanations that are useful for control than the ones that are deeply explanatory.

References

Hitchcock, Christopher, 1995. "Discussion: Salmon on Explanatory Relevance." *Philosophy of Science*, 62: 304-20.

Kitcher, Philip and Salmon, Wesley, 1989, *Scientific Explanation*. University of Minnesota Press, Minneapolis.

Newton, Isaac. 1726/1972. Isaac Newton's Philosophiae Naturalis Principia Mathematica. Edited by Alexandre Koyré and Bernard I. Cohen. Repr. Cambridge University Press.

Woodward, James 1997. "Explanation, Invariance and Intervention." *PSA 1996 2*: S-26-41, S26-S41.

Woodward, James. 2003. *Making Things Happen. A Theory of Causal Explanation*. Oxford University Press.

Piecewise Versus Total Support:

How to Deal with Background Information in Likelihood Arguments

Benjamin C. Jantzen

Virginia Tech

The probabilistic notion of *likelihood* offers a systematic means of assessing "the relative merits of rival hypotheses in the light of observational or experimental data that bear upon them."[1] In particular, likelihood allows one to adjudicate among competing hypotheses by way of a two-part principle:

Law of Likelihood (LL):[2]

(i) Evidence *E* supports hypothesis H_1 over H_2 just if $P(E|H_1) > P(E|H_2)$, where $P(E|H_i)$ is the likelihood of hypothesis H_i given evidence *E*.

(ii) The degree to which *E* supports H_1 over H_2 is measured by the *likelihood ratio*, $\Lambda = \frac{P(E|H_1)}{P(E|H_2)}$.

The claims sanctioned by LL are strictly comparative. The principle does not say what you should believe or to what degree you should believe it. Rather, the notion of 'supporting' one hypothesis over another is contrastive and perhaps best characterized as a relation of 'favoring'.[3] LL tells you how to determine the degree to which one hypothesis is favored over another on the basis of some evidence, E, and nothing more. Proponents of the principle are adamant that LL

[1] A. W. F. Edwards, *Likelihood* (Cambridge: Cambridge University Press, 1972) p. 1.

[2] I am using Elliot Sober's terminology here. The Law of Likelihood as I've presented it is to be distinguished from the weaker "Likelihood Principle," which in most formulations is equivalent to part (i) of LL. I caution the reader that both the terms "Law of Likelihood" and "Likelihood Principle" are used ambiguously in the philosophy of statistics and inductive inference literature.

[3] Elliott Sober, *Evidence and Evolution: The Logic Behind the Science* (Cambridge: Cambridge University Press, 2008).

cannot provide sufficient grounds for apportioning belief, only ranking hypotheses in a particular evidentiary context.

While LL has been defended at length as a general tool for both formal and informal reasoning about hypothesis ranking,[4] there remains an important ambiguity its application. Intuitively, we ought to make use of all available information when assessing the relative merits of two hypotheses, not just the particular piece of evidence *E* under consideration. Any additional information already in our possession prior to obtaining *E* is typically referred to as *background information*. LL does not, on the face of it, tell us how to deal with such information. Some, most prominently Elliott Sober[5], have argued that we ought to condition on this additional information when computing likelihoods. That is, if we denote the background information by *B*, then the likelihood ratio we should use is $\Lambda = \frac{P(E|H_1,B)}{P(E|H_2,B)}$. Taking this approach, however, means that Λ—and thus our judgments concerning rival hypotheses H_1 and H_2—will depend on exactly which information is taken to constitute background information, and which is considered evidence and thus part of *E*. Under Sober's interpretation, LL can be taken to yield different judgments for the

[4] See, for instance, Edwards, *Likelihood*; Ian Hacking, *Logic of Statistical Inference* (Cambridge: Cambridge University Press, 1965); Sober, *Evidence and Evolution: The Logic Behind the Science*.

[5] Elliott Sober, "The Design Argument," *God and Design*, ed. Neil Manson (New York, NY: Routledge, 2003) 27-54; Sober, *Evidence and Evolution: The Logic Behind the Science*; Elliott Sober, "Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads," *Philosophical Studies* 143 (2009): 63-90.

same data when the line between evidence and background information is moved. The use of LL is thus encumbered by a "line-drawing problem."[6]

This line-drawing problem also appears in a slightly different guise in the literature on statistical inference. In this more restricted context, the problem manifests as an apparent ambiguity in the likelihood function. Specifically, there appears to be no systematic way of deciding which random variables and model parameters should be included in the likelihood function, and no principled way of deciding on which side of the conditionalization bar these quantities belong if included.[7] As in the general case, the problem for the likelihoodist is to provide a principled division of propositions into background and evidence.

A variety of solutions have been proposed to both versions of the problem of background information, though not always in these terms. Some, e.g. Jonathan Weisberg,[8] attempt to provide a principled means of distinguishing evidence from background information. Others, e.g. Matthew Kotzen,[9] attempt to dissolve the problem by scrapping LL. In the context of statistical

[6] M. Kotzen, "Selection Biases in Likelihood Arguments," *The British journal for the philosophy of science* (2012).

[7] See M. J. Bayarri, M. H. DeGroot and J. B. Kadane, "What Is the Likelihood Function?," *Statistical Decision Theory and Related Topics Iv*, eds. Shanti S. Gupta and James O. Berger, vol. 1 (New York: Springer-Verlag, 1987) 3-27.

[8] Jonathan Weisberg, "Firing Squads and Fine-Tuning: Sober on the Design Argument," *British Journal for the Philosophy of Science* 56 (2005): 809-21.

[9] Kotzen, "Selection Biases in Likelihood Arguments."

inference, a common strategy is to disambiguate the likelihood function by fiat.[10] I argue that none of these strategies is well-motivated. Background information is only problematic when one fails to distinguish between two related questions: (i) Given that I know *B*, to what degree does the additional piece of evidence *E* support H_1 over H_2? and (ii) to what degree does all the evidence to hand—*B* and *E*—support H_1 over H_2? My aim is to demonstrate that, once these questions are distinguished the very same considerations that motivate the adoption of LL entail distinct answers to both questions, thus resolving any ambiguity over the treatment of background information. Note that I am emphatically not offering a defense of LL as a general inference procedure. Mine is the more modest goal of dissolving an apparent defect of LL using the resources to which proponents of the principle already assent.

To draw out the distinction relevant to eliminating the problem of background information, I will begin with a detailed example. I will then argue for an expression that represents the degree to which a particular piece of evidence supports one hypothesis over another in context, and then derive a related expression for the total support provided by all available evidence. Finally, I will show how these new expressions dissolve ambiguities in the treatment of background information by applying them to the so-called 'fine-tuning argument'.

I. ILLUSTRATING THE PROBLEM

[10] See, e.g., Jason Grossman, "The Likelihood Principle," *Philosophy of Statistics*, eds. Malcolm R. Forster and Prasanta S. Bandyopadhyay (Oxford, UK; Burlington, MA: North-Holland, 2011).

To draw out the distinction which I claim obviates the problem of background information, it will help to have a concrete example in mind. To avoid pre-conceived interpretations, I will intentionally eschew standard examples, at least at the outset. So rather than treat of fish or firing squads, I'll consider carnivals.

Suppose that Albert finds himself on the midway of an old-fashioned carnival. He decides to play one of the games—the one where contestants try to toss a ball into a milk-can. Albert is savvy about carnival games; he knows they are often rigged. In a fair game, there is a 50% chance of winning a prize. But when no authorities are around, there is an appreciable chance that the carnie running the game will hand him a ball too big to fit in the can, making it impossible to win. On the other hand, if there happens to be a police officer in sight the game is likely to be rigged in Albert's favor—the carnies want the police to think the games are fair, so they arrange to let people win when the authorities are present. A set of probabilities reflecting these facts is provided by the joint distribution of Table 1.

Table 1.

	P = police present		P = police absent	
	G = fair	G = rigged	G = fair	G = rigged
O = lose	1/20	1/20	1/10	11/20
O = win	1/20	1/10	1/10	0

Knowing all of the probabilities in Table 1, Albert puts his money down, and promptly tosses a ball into the can. Given that he has just won, what can Albert conclude about the game? Specifically, does he now have grounds to favor the hypothesis that the game is fair over the hypothesis that it is rigged? According to LL, Albert needs to compare two probabilities: the probability that he would win given that the game is fair, P(win| fair) and the probability that he would win given that the game is rigged P(win| rigged). Since P(win | fair) = 1/2 > P(win | rigged) = 1/7, LL asserts that Albert's success in the game supports the hypothesis of a fair game—Albert has reason to think that he has played a fair game.

But suppose that, before he tosses the ball, Albert notices a police officer standing near the booth. What can be said in light of this additional information? Here is where different interpretations of LL begin to diverge. According to Sober's approach, we must recognize two sorts of propositions: evidence and background knowledge. Evidence is whatever fresh information we are currently considering when applying LL to distinguish among hypotheses. It appears to the left of the conditionalization bar when computing a likelihood. Background knowledge constitutes whatever we already know about the world, and is presumed to belong on the right side of the conditionalization bar. According to this view then, Albert should treat the fact of the police officer's presence as background knowledge and condition on this information. The relevant likelihoods are now P(win| fair, present) = 1/2 and P(win| rigged, present) = 2/3. With the additional information, he should now favor the hypothesis that the game is rigged—the background information has reversed our ordering on hypotheses.

That we should take all available information into account when comparing hypotheses is not especially controversial—most authors assume some sort of *principle of total evidence.*[11] What is controversial is how and whether 'evidence' should be distinguished from background information. It is not clear why Albert should treat the information that a police officer was present any differently than the information that he won the game. Albert might just have well have treated the observation of the police officer as the evidence, and conditioned instead on the fact that he won: $P(\text{present} \mid \text{rigged, win}) = 1 > P(\text{present} \mid \text{fair, win}) = 1/3$. In this way of accounting for all the information, LL still favors the hypothesis that the game is rigged, but does so to a much greater degree. Alternatively, Albert might have treated all the information at hand as 'evidence' and compared the following likelihoods: $P(\text{win, present} \mid \text{fair}) = 1/6 > P(\text{win, present} \mid \text{rigged}) = 1/7$. Taking this approach once again inverts the ordering of hypotheses, and favors the hypothesis that the game was fair. It might appear then that LL must be modified in order to provide a principled means of discriminating background information from evidence. However, no such modification is required—a careful interpretation of LL as it stands obviates the question of evidence versus background information.

II. THE PIECEWISE IMPACT OF EVIDENCE

To resolve the ambiguity over background information, we need to distinguish between two questions: (i) to what degree does learning a particular fact in the context of an additional set of facts support a given hypothesis? and (ii) to what degree does learning a particular fact in conjunction with an additional set of facts support a given hypothesis? In terms of the midway

[11] Rudolph Carnap, "On the Application of Inductive Logic," *Philosophy and Phenomenological Research* 8, 1 (1947): 133-48.

example above, the distinction can be made as follows: (i) to what degree does winning the game having already learned that a police officer is present support the hypothesis that the game is fair? and (ii) to what degree does the full set of information at hand—that Albert has won the game and that a police officer was present—support the hypothesis that the game is fair?

To address question (i), we need to examine the piecewise introduction of evidence, taking care to note one important fact: learning the truth of a proposition (or the value of a random variable) is effectively an intervention that changes the background distribution describing the ways the world might be. To begin with, let's assume that we are given a full joint distribution reflecting all relevant aspects of the world and nothing else—there is nothing given that might qualify as either evidence or background information. For ease of exposition, I will further assume that this distribution is discrete, though nothing about my derivation hinges on this being the case.

Since all we have is the distribution and no information to sort out, LL can be applied unambiguously upon obtaining our first piece of evidence, I_1. According to LL, the degree to which this information supports hypothesis H_1 over H_2 is given by the likelihood ratio $\Lambda(I_1) = P(I_1|H_1)/P(I_1|H_2)$. Furthermore, on learning that I_1 is the case, the space of possible events has been reduced—acquiring information requires us to update the background distribution with which we started. Specifically, the probability of I_1 being the case must now be unity, irrespective of the value it had prior to learning this outcome. One way to represent the change is to construct a new event space by simply removing all the events incompatible with the fact that I_1 is the case while preserving the relative measure on all remaining events. That is, the new distribution $P_1(\alpha)$, where α is any event in the original event space compatible with I_1, is

obtained from the old distribution by the following relation: $P_1(\alpha) = P(\alpha|I_1)$.[12] In the midway example, for instance, when Albert learned that a police officer was present he should have replaced the original distribution of Table 1 with that of Table 2.

Table 2.

	P = police present	
	G = fair	G = rigged
O = lose	1/5	1/5
O = win	1/5	2/5

Once we realize that we are working with a new distribution, there is no need to draw a line between background information and evidence—our prior information is reflected in the new distribution. When additional evidence, I_2, is acquired, we need only appeal to LL just as we did at the outset. This time, however, we are assessing likelihoods with respect to the currently applicable distribution $P_1(\alpha)$. So the evidence I_2, if we take LL seriously, supports H_1 over H_2 just if $P_1(I_2|H_1) > P_1(I_2|H_2)$ and does so to a degree $\Lambda(\mathrm{I}_2) = P_1(I_2|H_1)/P_1(I_2|H_2)$. In terms of the original joint distribution, we can express this likelihood ratio as $\Lambda(\mathrm{I}_2) = P(I_2|I_1, H_1)/P(I_2|I_1, H_2)$.

[12] This is simply the updating procedure recommended by Bayesian epistemology. It is invoked here without any commitment to the subjective or objective status of priors.

As before, when we learn I_2, we must update our distribution to reflect this restriction of the possibilities. This new distribution $P_2(\beta)$ is obtained from the old distribution in the same way as above: $P_2(\beta) = P_1(\beta|I_2) = P(\beta|I_2, I_1)$. This is easy to generalize for an indefinite sequence of evidence: once we've learned I_1, I_2, …, I_{n-1}, we should compute the likelihoods involving a new piece of evidence I_n using the distribution $P_{n-1}(\gamma) = P(\gamma|I_{n-1}, \ldots, I_1)$. The new piece of information I_n introduced in the context of prior information I_1, I_2,…, I_{n-1} supports H_1 over H_2 just if $P(I_n|I_{n-1}, \ldots, I_1, H_1) > P(I_n|I_{n-1}, \ldots, I_1, H_2)$ and does so to the degree

(1) $\Lambda(I_n) = \frac{P(I_n|I_{n-1},\ldots,I_1,H_1)}{P(I_n|I_{n-1},\ldots,I_1,H_2)}$.

The point is that whenever we acquire a piece of information we can apply LL without modification, but must do so using a distribution that reflects all of the facts already in evidence. Put this way, there is no ambiguity in using LL—we always compute a straightforward likelihood. However, when this likelihood is expressed in terms of the original joint distribution with which we started, each successive likelihood is conditioned on the previous facts. So by applying LL and taking care to note the way in which the acquisition of information forces a change in distribution, we have found that in order to determine the relative support of one hypothesis over another provided some particular piece of evidence, we must use likelihoods conditioned on all previously acquired facts.

Thus far, it may seem that I have been arguing for Sober's interpretation of LL. However, Sober seems to view the likelihood ratio (1) as representing the overall degree to which H_1 is supported over H_2 once I_n is obtained. I have been urging that, if we take LL at face value, this is *not* how

we should interpret this expression. At every stage in the above derivation, we were applying LL to determine the degree to which a particular piece of evidence supported one hypothesis over another. Other information was relevant, but only in determining the epistemic context in which this degree of support was determined. I am suggesting that Sober has the right expression but gives it in answer to the wrong question—in what follows, I'll show that LL leads us to a very different expression for the degree of support for H_1 over H_2 provided by the totality of evidence.

III. TOTAL SUPPORT

There are two ways to argue for an expression of the likelihood ratio pertaining to the totality of available evidence. In one approach, we could take the expression given in (1) for the degree to which a particular piece of evidence supports H_1 over H_2 and couple this with a function for combining likelihood ratios—a function measuring the overall degree to which two pieces of evidence support H_1 over H_2. Strictly speaking, this means adding to LL since the principle does not provide such a rule. However, there are some reasonable constraints we can put on such a function without begging the question concerning background information. For starters, whatever function f we choose should itself yield a likelihood ratio, meaning that it must map pairs of likelihoods to the interval $[0, \infty)$. Furthermore, if either likelihood in the combination is zero—implying that one hypothesis has been entirely ruled out—then the joint likelihood should also be zero. The function should be symmetric since it ought not to matter in what order we give the likelihoods to be combined, and it should be an increasing function of both arguments. An obvious choice satisfying all of these constraints is simply the product of the component likelihoods. That is, given Λ_1 and Λ_2, the combined likelihood is given by $f(\Lambda_1, \Lambda_2) = \Lambda_1\Lambda_2$. With this rule for combining likelihoods, we can use the results of the last section to derive an

expression for the overall degree to which the facts I_1, I_2, …, I_n support one hypothesis over another, assuming they were learned in sequence:

(2) $$\Lambda(I_1, I_2, \ldots, I_n) = \Lambda(I_1)\Lambda(I_2)\cdots\Lambda(I_n) = \frac{P(I_1|H_1)P(I_2|H_1,I_1)\cdots P(I_n|H_1,I_1,\ldots,I_{n-1})}{P(I_1|H_2)P(I_2|H_2,I_1)\cdots P(I_n|H_2,I_1,\ldots,I_{n-1})}$$

Using nothing but the rules of probability, the right hand side of equation (2) can be written much more compactly to give the following expression for the total support of the facts I_1, I_2, …, I_n:

(3) $$\Lambda(I_1, I_2, \ldots, I_n) = \frac{P(I_1,\ldots,I_n|H_1)}{P(I_1,\ldots,I_n|H_2)}$$

Of course, the right-hand side of equation (3) is just the expression we would have gotten by applying LL to the proposition I_1^I_2^…^I_n with respect to the initial joint distribution—in a straightforward reading, it is just the total support for H_1 over H_2 provided by the conjunction of all available evidence.

The form of Equation (3) suggests that it might have been derived more directly by appealing to LL without worrying about how to determine the contextual support provided by each piece of information or introducing a way to combine these (thus justifying my claim that we need not modify LL). All we had to do was note that, if we let $E = I_1$^I_2^ ... ^I_n, then LL immediately yields (3). From (3) we could then deduce (2) just from the rules of the probability calculus. Once we identified the factors of the right-hand side of Equation (2) with individual likelihood ratios, we could have used this fact to justify a rule for combining likelihoods. In fact, this is what A. F. Edwards does, at least in the special case of independent evidence, in his development of the likelihood framework.[13] Viewed from this perspective, Equation (3) is implicit in LL.

[13] Edwards, *Likelihood*.

Whichever approach we take to justifying this rule for assessing total support, we are led to the following amplified form of LL:

Amplified Law of Likelihoods (ALL):

(i) If it is already known to be that case that I_1^I_2^…^ I_n, then learning evidence E supports hypothesis H_1 over H_2 just if $P(E|H_1, I_1, I_2, \ldots, I_n) > P(E|H_2, I_1, I_2, \ldots, I_n)$, where $P(E|H_i, I_1, I_2, \ldots, I_n)$ is the likelihood of hypothesis H_i given evidence E in the context of I_1^I_2^…^ I_n.

(ii) The degree to which E supports H_1 over H_2 in the context of I_1^I_2^…^ I_n is measured by the likelihood ratio $\Lambda = \frac{P(E|H_1, I_1, I_2, \ldots, I_n)}{P(E|H_2, I_1, I_2, \ldots, I_n)}$.

(iii) The total evidence E^ I_1^I_2^…^ I_n supports hypothesis H_1 over H_2 just if $P(E, I_1, I_2, \ldots, I_n|H_1) > P(E, I_1, I_2, \ldots, I_n|H_2)$.

(iv) The degree to which the total evidence E^ I_1^I_2^…^ I_n supports H_1 over H_2 is measured by the likelihood ratio $\Lambda = \frac{P(E, I_1, I_2, \ldots, I_n|H_1)}{P(E, I_1, I_2, \ldots, I_n|H_2)}$.

With ALL, we can answer the questions posed above concerning the midway example. The information that Albert has won the game, acquired after learning that a police officer is present, supports the hypothesis that the game is rigged because $P(\text{win}|\text{present, rigged}) > P(\text{win}|\text{present, fair})$. According to ALL (ii), this information favors the rigged hypothesis over its rival to a degree $\Lambda = \frac{P(\text{win}|\text{present, rigged})}{P(\text{win}|\text{present, fair})} = \frac{\frac{2}{3}}{\frac{1}{2}} = \frac{4}{3}$. This one piece of information, in the context of previously established information about the presence of police officers, tends to favor the

hypothesis of a rigged game. However, the aggregate information—that a police officer is present and Albert has won the game—favors the hypothesis that the game is fair. This follows from ALL (iii) and (iv) since $\frac{P(\text{win, present}|\text{ fair})}{P(\text{win, present}|\text{ rigged})} = \frac{\frac{1}{6}}{\frac{1}{7}} = \frac{7}{6}$. This looks like a contradiction until we realize that the first piece of information obtained—that the police officer is present—strongly favored the hypothesis that the game is fair: $\frac{P(\text{present}|\text{fair})}{P(\text{present}|\text{rigged})} = \frac{14}{9}$. The upshot is that the aggregate effect of the totality of evidence can differ from the piecewise impact of each bit of evidence. Rather than being a contradiction, this is precisely how one would expect these two distinct measures to relate—the total support for the fair hypothesis is simply the product of the contextual likelihood ratios for each piece of evidence.[14]

IV. KICKING AWAY THE FULL DISTRIBUTION LADDER

In the preceding arguments, I made extensive use of full probability distributions. This appears problematic since the appealing feature of the likelihood approach—and that which sets it apart from Bayesianism—is its disregard for prior probabilities. However, I claim that the likelihoodist who thinks that prior probabilities are often absent or unattainable might nonetheless justify LL or ALL. To see how, let's reconsider the case in which we start with a full prior distribution $P(\alpha)$, and then obtain evidence I_1. Once we acquire the evidence, we should update the probabilities assigned to H_1 and H_2 by setting each new probability equal to the corresponding conditional probability assigned by the original distribution:

[14] It should be noted that, while the order in which information is learned determines the degree to which each additional piece of information favors one hypothesis over another, order is irrelevant when considering the overall support conferred by the totality of evidence.

$$P\left(H_1\middle|I_1\right)=P\left(I_1\middle|H_1\right)\frac{P\left(H_1\right)}{P\left(I_1\right)},\ P\left(H_2\middle|I_1\right)\quad P\left(I_1\middle|H_2\right)\frac{P\left(H_2\right)}{P\left(I_1\right)}$$

What can we now say about the degree to which I_1 favors H_1 over H_2? One way we might understand this question is in terms of a hypothetical. Suppose that either H_1 or H_2 is true. Then the initial odds in favor of H_1 are simply $P(H_1|H_1\vee H_2)/P(\sim H_1|H_1\vee H_2) = P(H_1)/P(H_2)$. How does the new information change the odds in favor of H_1? In this case, the posterior odds are given by:

$$(4)\qquad \frac{P\left(H_1\middle|I_1,H_1\vee H_2\right)}{P\left(\sim H_1\middle|I_1,H_1\vee H_2\right)}=\frac{P\left(I_1\middle|H_1\right)P\left(H_1\right)}{P\left(I_1\middle|H_2\right)P\left(H_2\right)}=\Lambda\left(I_1\right)\frac{P\left(H_1\right)}{P\left(H_2\right)}$$

The right-hand equality in Equation (4) indicates that all of the work to shift the posterior odds up or down relative to our prior odds is being done by the likelihood ratio, $\Lambda(I_1)$. In other words, the change in posterior odds is a function of $\Lambda(I_1)$. To put it still another way, the effect of I_1 on the odds is entirely determined by $\Lambda(I_1)$. This fact motivates adopting the likelihood function as a measure of relative support. While the likelihood ratio cannot tell us which posterior probability is higher, it can tell us how the odds shift in favor of one hypothesis or the other, assuming that one or the other is right. Furthermore, it does so whether or not we know the prior probabilities. In this sense, LL is a general guide to differential support, and in those cases in which we have no objective basis for assigning priors, the likelihoodist claims it is our only guide.

By considering effects on posterior odds, we can motivate ALL in much the same way as LL. As before, the full distribution (if we knew it to begin with) after learning I_1 would be given by $P_1(\alpha) = P(\alpha|I_1)$. If we now learn that I_2 is the case, then we must change our posterior odds in favor of H_1 over H_2 to the following:

$$(5) \qquad \frac{P_1\left(H_1 \middle| I_2, H_1 \vee H_2\right)}{P_1\left(H_2 \middle| I_2, H_1 \vee H_2\right)} = \frac{P_1\left(I_2 \middle| H_1\right)P_1\left(H_1\right)}{P_1\left(I_2 \middle| H_2\right)P_1\left(H_2\right)} = \Lambda\left(I_2\right)\frac{P_1\left(H_1\right)}{P_1\left(H_2\right)}$$

Once again, it is the likelihood function that increases (or decreases) the posterior over the prior odds. This time, however, it is in the context of the new distribution $P_1(\alpha)$, a distribution reflecting prior knowledge of I_1. If the motivation offered for LL in the first place is compelling, then it seems we must also accept ALL (i) and (ii)—the relative support for H_1 over H_2 conferred by the new piece of evidence I_2 after already learning I_1 is indicated by the likelihood function, $\Lambda(I_2) = P_1(I_2|H_1)/P_1(I_2|H_2) = P(I_2|I_1, H_1)/P(I_2|I_1,H_2)$. But what about the overall support for H_1 over H_2 given our epistemic starting point? How should our posterior odds have changed relative to our initial odds as a result of learning I_1 and I_2? We can rewrite the right-hand side of Equation (5) as follows:

$$(6) \qquad \begin{aligned} \frac{P_1\left(I_2 \middle| H_1\right)P_1\left(H_1\right)}{P_1\left(I_2 \middle| H_2\right)P_1\left(H_2\right)} &= \frac{P\left(I_2 \middle| H_1, I_1\right)P\left(H_1 \middle| I_1\right)}{P\left(I_2 \middle| H_2, I_1\right)P\left(H_2 \middle| I_1\right)} \\ &= \frac{P\left(I_1, I_2 \middle| H_1\right)P\left(H_1\right)}{P\left(I_1, I_2 \middle| H_2\right)P\left(H_2\right)} \\ &= \Lambda\left(I_1, I_2\right)\frac{P\left(H_1\right)}{P\left(H_2\right)} \end{aligned}$$

Written this way, we can see that the combined likelihood function $\Lambda(I_1, I_2)$ determines the change in odds relative to what they were before learning anything at all. So once again, we can kick away the ladder of the full distribution. If we did know the distribution, then learning I_1 and I_2 would change the odds in favor of H_1 over H_2 by an amount given by the likelihood ratio. Since this is true irrespective of what the priors are, we can always take the likelihood alone to indicate differential support, in this case the degree of support for H_1 over H_2 conferred by the totality of evidence.

Of course, one might object to my interpretation of what it is to favor one hypothesis over another. Instead, one might attempt to prove LL(i) from other premises[15] and take the quantitative measure of contrastive support given by LL (ii) to be a postulate that stands or falls with how well the results coincide with our intuitions.[16] ALL could then be motivated by the second line of argument I suggested in the previous section: treat all information as evidence and note that the resulting likelihood ratio factors into a product of likelihoods, each of which can be consistently interpreted as corresponding to the impact of a single piece of information.

The point is that insofar as LL is well-motivated, so too is ALL. My use of full distributions above was strictly heuristic. Once we've seen what role the likelihood function plays and which likelihood function is relevant to which question, we can ignore the full distribution. Of course, the proponent of LL or ALL can only claim to be free of worrisome priors if conditional probabilities can be taken as primitive.[17] It is not my task here to defend that claim and thus rescue likelihoodism from the charge of subjectivity. My more modest assertion is simply that if we have grounds to take LL seriously, then we should really embrace ALL. Once we've done so, it becomes clear that whatever problems likelihoodism has, line-drawing isn't one of them.

V. BUT WHAT IS THE LIKELIHOOD FUNCTION?

[15] See Grossman, "The Likelihood Principle."

[16] See Sober, *Evidence and Evolution: The Logic Behind the Science*.

[17] For a defense of taking conditional probabilities as primitives, see ibid.

My arguments so far have concerned the problem of background information as it appears in the literature on LL in its broadest epistemic use. As I mentioned above, the same problem arises in the more restricted context of statistical inference. Addressing this narrower community, Bayarri, De Groot, and Kadane famously asked, "What is the likelihood function?"[18] To illustrate the ambiguity in answering that question, the authors consider a case analogous to one in which, with respect to each possible value of some discrete parameter θ characterizing a statistical model, a random variable X has a conditional probability distribution $P(x|\theta)$.[19] Furthermore, it is not the random variable X that is observed, but rather some other random variable Y for which $P(y|x, \theta)$. The authors then ask, "What is the [likelihood function] in this problem?".[20] They claim that there are three candidates, $P(y|\theta)$, $P(x, y|\theta)$, and $P(y|x, \theta)$, and that a "subjective judgment must be made in order to decide which of the functions…to use in a given problem."[21] The thesis I've been defending is that this is simply false. The question has two parts: (i) which random variables and parameters are to be included in the likelihood function, and (ii) which side of the conditionalization bar each belongs on. The answer to both parts, according to ALL, depends on two things: what hypotheses we wish to consider and whether we wish to assess the impact of a particular piece of data in context or the aggregate of all data. So, for instance, suppose we wish to ask about hypotheses concerning the value of θ in light of the only piece of

[18] Bayarri, DeGroot and Kadane, "What Is the Likelihood Function?."

[19] The original example was stated in terms of probability densities since θ typically takes a continuum of values. To keep the discussion consistent, I've assumed that θ is discrete, and thus the distributions in question are discrete as well.

[20] Bayarri, DeGroot and Kadane, "What Is the Likelihood Function?," at p. 6.

[21] Ibid., 6.

evidence available, namely a value y of Y. Then the relevant likelihood function must have the form $P(y|\theta)$. If on the other hand, we wanted to consider finer-grained hypotheses concerning both the value of θ and the unobserved random variable X, then we would have functions of the form $P(y|x, \theta)$. Under no circumstances would ALL entail the use of a likelihood function of the form $P(x|\ldots)$ unless a value of the random variable X was observed (or otherwise learned) and thus added to our store of facts. Suppose X and Y were both observed and we wish to know the relative support given to hypotheses about θ. Then our likelihood functions would look like $P(x,y|\theta)$. Suppose instead, we learned the value of Y and then the value of X and wish to know what impact learning $X = x$ has given what we already know about Y. Then the likelihood functions would have the form $P(x|y, \theta)$. I'm belaboring the point, but I want to make it clear that ALL unambiguously selects a set of variables and parameter values and distributes these around the conditionalization bar. There are many further objections raised by Bayarri et al to the use of LL as a statistical inference method, in particular problems with prediction. However, many of these objections conflate LL (or ALL) with the method of maximum likelihood estimation (MLE). A discussion of the relation of MLE to ALL is beyond the scope of this paper, and so too are the remaining objections to likelihoodism. It suffices here to note that there is no ambiguity in factoring the likelihood function as far as ALL is concerned. The principle may not be right, but it is unambiguous.[22]

[22] The authors might object that in my initial discussion of ALL, I used a full distribution which dictates all the relevant quantities and so implicitly settles the question of which likelihood function to use. But as I argued in the last section, a full distribution is unnecessary for motivating ALL. Rather, in the likelihoodist view, specifying a question of interest specifies a

VI. FISH, FIRING SQUADS, AND FINE-TUNING

The question of how to handle background information is especially pressing in the context of the *fine-tuning argument* (FTA). The FTA attempts to establish the existence of a cosmic designer by noting that various physical constants have values within a narrow range amenable to the occurrence of carbon-based life—the laws appear 'fine-tuned' for life. For instance, had the 7.65 MeV energy level of the C^{12} nucleus been slightly lower or higher, then the process that produces carbon and the other heavy elements essential to life in the interior of stars would not have occurred.[23] Denote by *E* the observation that many constants occurring in physical laws take values within a comparatively narrow range that permits life to exist, and consider the following two hypotheses:

H_C: The relevant physical constants acquired their values by chance.

H_D: The relevant physical constants acquired their values by design.

The FTA is usually presented as a likelihood argument. If we appeal to LP and note that $P(E|H_D) > P(E|H_C)$, then we must conclude that the evidence favors design over chance.

A prominent objection to the fine-tuning argument notes that we have left out an important piece of information: all knowledge concerning physical constants has been acquired by carbon-based

---

likelihood ratio which in turn constrains what full distributions the Bayesian (or anyone else committed to using full distributions) may consider.

[23] John D. Barrow and Frank J. Tipler, *The Anthropic Cosmological Principle* (New York: Oxford University Press, 1986) pp. 252-53.

life forms.[24] Call this fact *I*. We must account for all available background information—so the objection goes—and so we must condition our likelihoods on *I*. However, since *I* entails *E*, both hypotheses have the same likelihood given the evidence: $P(E|H_D, I) = P(E|H_C, I) = 1$. Thus, the evidence cannot favor design over chance (or any other hypothesis for that matter). This objection, however, conflates the two questions with which we began and emphasizes the need for the clarification provided by ALL.

To motivate an analysis of the FTA in terms of ALL, it will help to first consider a pair of structurally similar examples endemic in the literature. The first of these, due originally to Sir Arthur Eddington,[25] asks us to think about fishing. Suppose we are confronted with the following observation:

E_f: All 10 of the fish caught in the lake today were longer than 10 inches.

For the sake of simplicity, suppose that we consider only two hypotheses that might account for this evidence:

H_{100}: All of the fish in the lake are longer than 10 inches.

H_{50}: Half of the fish in the lake are longer 10 inches.

If this was all the information we had, LP would urge us to favor H_{100} since $P(E_f|H_{100}) \gg P(E_f|H_{50})$. However, suppose we had some additional information:

$I_{>10}$: The net used has holes 10 inches wide.

---

[24] Sober, "The Design Argument."; Sober, "Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads."

[25] A. Eddington, *The Philosophy of Physical Science* (Cambridge: Cambridge University Press, 1947).

This new information $I_{>10}$ entails E_f. Thus, if we account for this new information by conditioning on it as Sober would urge, we find that the evidence fails to distinguish between the hypotheses at all: $P(E_f|H_{100}, I_{>10}) = P(E_f|H_{50}, I_{>10}) = 1$. According to Sober, this constitutes an *Observation Selection Effect* (OSE) because the method by which the observation was obtained biased the outcome. One is faced with an OSE whenever accounting for the process by which an observation was made alters the likelihoods that determine the degree to which the observation favors one hypothesis over another. In this case, the effect is extreme.

The picture changes dramatically when we analyze this scenario using ALL. It becomes immediately obvious that the likelihoods being compared— $P(E_f|H_{100}, I_{>10})$ and $P(E_f|H_{50}, I_{>10})$—represent only the degree to which learning about the day's catch supports either H_{100} or H_{50} in the context of information about the net used. These do not represent the degree to which the aggregate evidence supports one or the other hypothesis. It is true that learning E after learning what net was used fails to further discriminate between H_{100} and H_{50}. But learning $I_{>10}$ may have already discriminated between the two, and thus, according to ALL, the aggregate information might also discriminate between the two hypotheses.

To illustrate the point, consider the joint distribution in Table 3. I've added a proposition, $I_{>0}$, which is the claim that the net used had very tiny holes capable of catching the smallest fish. With this additional possibility added, the probabilities given are compatible with all of the facts above. In particular, $P(E_f|H_{100}) = 1 \gg P(E_f|H_{50}) = .003$ and $P(E_f|H_{100}, I_{>10}) = P(E_f|H_{50}, I_{>10}) = 1$.

Table 3.

	H_{100}		H_{50}	
	$I_{>0}$	$I_{>10}$	$I_{>0}$	$I_{>10}$
E_f	.001	.002	.001	.002
$\neg E_f$	0	0	.994	0

However, we can see that learning $I_{>10}$ at the outset strongly favored the hypothesis H_{100} since $P(I_{>10}|H_{100}) = 0.67 \gg P(I_{>10}|H_{50}) = 0.002$. Likewise, according to ALL (iv), the aggregate information overwhelmingly favors H_{100} over H_{50} to a degree given by $\Lambda = P(E_f, I_{>10}|H_{100})/P(E_f, I_{>10}|H_{50}) = 334$. This conclusion is not surprising given the details of the example. The distribution given in Table 3 is plausible in that those who frequently fish a particular lake are more likely to use nets with large holes if the lake contains mostly large fish—they may not know the distribution of fish in the lake, but they know what works. Whatever story one might tell to account for the particular probabilities in this case, the upshot is that if an OSE renders a particular observation irrelevant in a particular context it is still possible for the aggregate information to discriminate between hypotheses.

While Eddington's fishing example illustrates the way in which previously acquired information can deprive subsequent evidence of relevance, there is another example in the literature more

closely analogous to the fine-tuning case.[26] This scenario involves firing squads. We are asked to imagine that a firing squad staffed by twelve expert marksmen takes aim at the prisoner to be executed. Each marksman fires twelve times when given the signal. When the smoke clears, we discover that the prisoner is still unharmed. Call the fact of this surprising survival E_s. In this case, we are interested in what the prisoner can infer from E_s concerning the following two hypotheses:

H_{con}: The marksmen conspired at time t_1 to spare the prisoner's life when they fired at t_2.

H_{miss}: The marksmen decided at time t_1 to shoot the prisoner when they fired at t_2 but missed by chance.

At first we might think that the prisoner has ample reason to favor H_{con} over H_{miss} since, given that these are expert marksmen, $P(E_s|H_{con}) \gg P(E_s|H_{miss})$. However, in making his analysis the prisoner left out some pertinent information about the manner in which the observation of E_s was made:

I_O: At t_3 the prisoner made the observation that he is still alive.

According to those who would single out background information, we must incorporate I_O into the likelihoods by conditioning. In this view, the prisoner suffers from an OSE and cannot distinguish between the two hypotheses at all since $P(E_s|H_{con}, I_O) = P(E_s|H_{miss}, I_O) = 1$. Because I_O entails E_s, so the argument goes, learning E_s can tell the prisoner nothing about which

[26] The scenario was introduced in John Leslie, *Universes* (London: Routledge, 1989). and elaborated in Richard Swinburne, "Arguments from the Fine-Tuning of the Universe," *Physical Cosmology and Philosophy*, ed. J. Leslie (New York: MacMillan, 1990) 160-79.

hypothesis to favor. Thus, the prisoner in the grip of a strong OSE cannot reasonably conclude there was a conspiracy to save his life.

At this point, the tight analogy with the FTA should be clear. The prisoner stands in for us carbon-based life forms. While the prisoner is attempting to assess whether design or chance is responsible for his survival, in the FTA we are attempting to infer design in the cosmos. In both cases, it has been objected that the observer suffers from an OSE that prevents discrimination between hypotheses. Supporters of the FTA invoke the firing-squad scenario because they think that our intuition strongly opposes the OSE objection—surely the prisoner can reasonably conclude that conspiracy is the better hypothesis. By analogy, they claim that we can conclude that an OSE is not a problem for the FTA.

In both cases, ALL tells us that the role of the OSE has been misinterpreted. It is true that, in the context of knowing that it was himself who made the observation, the prisoner learns nothing further by noting that he is alive. Likewise, it is the case that, knowing that all physics is done by carbon-based life forms, we learn nothing further by discovering that the constants of physical law are just right to sustain carbon-based life. Nonetheless, the aggregate information might still favor one hypothesis over the other. In the firing-squad scenario, it is eminently plausible that $P(E_s, I_O|H_{con}) \gg P(E_s, I_O|H_{miss})$. In the case of fine-tuning, it may be that $P(E, I|H_D) > P(E, I|H_C)$. This will be the case if $P(I|H_D) > P(I|H_C)$. I certainly do not wish to argue that this is in fact the case—there seem to be insurmountable difficulties in providing a well-defined

measure corresponding to $P(I|H_D)$.[27] My point is just that, when one distinguishes between contextual and total support, the presence of an OSE does not prove fatal to design arguments in either the firing-squad or FTA case.

VII. CONCLUSION

Insofar as one is inclined to accept LL as a framework for inference, no modification is necessary in order to deal with background information—unpacking LL leads to ALL. The interpretive key is the discrimination of two questions, one concerning the immediate support provided by a piece of evidence in context and one concerning the overall support provided by the total set of evidence. Looked at in this way, it becomes clear that objections based on observer bias are not necessarily fatal to the FTA. It is true that we, as carbon-based life-forms, cannot use the fact that some physical constants are just right for the existence of carbon-based life to discriminate between design hypotheses and their rivals. However, it may be the case that the aggregate evidence (including the fact of our existence) might permit such discrimination. Whether this is the case must be settled on other grounds.

[27] It is not clear that the question of fine-tuning is even well-posed. There is reason to reject the strong metaphysical assumptions necessary to make the possibility of different 'constants' in the laws of nature meaningful or to entertain the existence of processes—whether physical or divine—that determined those constants in the past.

Critical Subjects: Participatory Research needs to Make Room for Debate

Inkeri.Koskinen@helsinki.fi

PSA 2012 Biennial Meeting
November 15, Session 1
Contributed Papers: Issues for Practice in Medicine and Anthropology

Abstract: *Participatory research in anthropology attempts to turn informants into collaborators, even colleagues. Researchers generally accept the idea of different knowledge systems, and the practice of avoiding critical appraisal of alien knowledge systems, common in ethnography, is continued within participatory research. However, if the aim of participatory research is to turn informants into collaborators, or ideally colleagues, the ethical imperative of offering constructive criticism to colleagues should apply to them, too, even if they are seen as representing different knowledge systems than the researchers. Avoiding appraisal of alien knowledge systems is problematic when the knowledge systems of the researcher and the researched are in constant contact.*[1]

[1] I would like to thank the staff of Sámi Allaskuvla, especially Nils Oskal, for their generosity during my short visit in 2011, and Jelena Porsanger for her kind answers to my e-mails. Irja Seurujärvi-Kari gave important help in organizing the visit, for which I owe her many thanks. I am also grateful for having had the opportunity to hear some presentations and discuss with some staff members at The First Nations University of Canada during the IAPL conference in 2010. Earlier versions of this paper have been presented in the Conference on the Philosophy of the Social Sciences in Copenhagen on 25.-26.8.2011, in the Finnish Centre of Excellence in the Philosophy of the Social Sciences' seminar in Helsinki on 16.1.2012 and in the VII Ethnology Days on 16.-17.3.2012 in Jyväskylä.

1. Introduction

The last few decades have witnessed the proliferation of different kinds of participatory, collaborative, ethnocritical and co-operative research methods in many disciplines. What the greater part of these methods have in common, is the attempt to change the relationship between the researcher and the researched from one between subject and object to one between subject and subject (Smith 1997, 178), and to turn informants, or local non-academic interest groups, into collaborators, even colleagues. The main focus here is on the use of these methods in ethnographic research. For the purposes of this paper I will call these forms of "academic engagement with outside communities" (Petras, Porpora 1993, 107) *participatory research*. Participatory research is mostly very down-to-earth and deals with questions and social problems that have weight in the daily lives of the communities the researchers work with. The reasons given for the adoption of such methods are mainly ethical, and when also epistemic grounds for the need of participatory research are discussed, they tend to be strongly attached to discussions concerning power inequalities: The position of a researcher is seen as a position of power, and researchers should be aware of the power structures they might consolidate by their work. The importance of the research subject's own knowledge is emphasized. Researchers should relinquish the idea of holding knowledge that would be privileged compared to that of the researched, who typically have a much lower social status than the researcher. Oppressed groups can, according to this vein of thought, be epistemically privileged, and researchers can benefit from their knowledge. (Finnis 2004, Hall 2005, Kurelek 1992, Park 2006, Wylie 2003.) Theoretical discussions about participatory research focus strongly on ethical issues. This paper takes an epistemic point of view, though the argument is nevertheless partly ethical.

Ethnographers generally accept the idea of there being different knowledge systems: people around the world have differing criteria for what is considered as a good argument and what is accepted as knowledge, or an acceptable way of producing knowledge claims. According to a widespread interpretation, these criteria are seen as stemming from - and as an integral part of – a conceptual framework. And the conceptual frameworks have especially earlier on been understood as chiming with *cultures*, understood as holistic systems that ethnographers could interpret. In trying to avoid ethnocentrism, ethnographers have developed research practices in which hasty comparisons between statements made in different knowledge systems are avoided: comparison as well as adjudication can be meaningful only when the position of the statement within its proper framework is understood. Shortly, many ethnographers *avoid appraisal* of alien knowledge systems.

The practice of avoiding appraisal is often linked to some form of relativism. As Mark Risjord has noted (1998), relativism does not necessarily lead to the impossibility of criticism, or avoiding appraisal of alien knowledge systems. But as shall be shown, avoiding appraisal follows easily from methodological conceptual relativism. It can be discerned also from recent ethnographic research inspired by postmodern epistemic relativism – notably, participatory research. The ethical and power-related arguments given for the adoption of participatory methods do not seem to lead to the abandoning of the practice of avoiding appraisal. Rather, researchers are encouraged to adopt a strictly positive attitude towards the local knowledge of the communities they are studying (Finnis 2004). The main goals of this kind of research are often social change, emancipation and 'giving back to the communities'. Accordingly, it seems much more interesting to use local knowledge in research when possible, than to critically appraise it. The ideal situation would be one where local knowledge and

"western" academic knowledge could be seamlessly incorporated, and the informant would thus turn into a co-author and effectively a colleague. But, as I will argue, postmodern epistemic relativism does not offer tools to analyse and deal with situations where the local and academic knowledge systems clash.

Avoiding appraisal is practicable only as long as the research subjects go along with it and the different knowledge systems stay at least somewhat apart. This is not always the case. The typical research subjects of cultural research have become more critical of their role as research subjects than they used to be in the heyday of 20th century anthropology. This change is by no means limited to cultural research; the general public's attitudes towards science and research have become more distrustful than it used to be (Carrier & Weingart 2009). In cultural research this change nevertheless has some unique features. An extreme demonstration of how research subjects have become critical of their role is the birth of a new and heterogeneous discipline called indigenous studies. Indigenous researchers wish to base their research methods on their own peoples' knowledge systems, which they hold to be different from the "western" ones (Tuhiwai Smith 1999). When such critical subjects enter academia, it becomes impractical to avoid appraisal of different knowledge systems, and it seems to become ethically questionable, too: constructive criticism is a researchers' due, and giving it is an obligation. Constructive criticism and avoiding appraisal are not compatible, so the practice of avoiding appraisal of different knowledge systems is ethically problematic when the alleged different knowledge systems enter academia. Moreover, if the aim of participatory research is to incorporate local knowledge with academic knowledge and turn informants into collaborators, and effectively even colleagues, the same ethical imperative applies to them, too, even if they are seen as

representing different knowledge systems. Their knowledge should be critically appraised.

To argue for this position, I shall start by discussing the practice of avoiding appraisal of different knowledge systems in ethnographic research. Then I shortly describe the development that has lead to the establishment of indigenous studies, and the general aims of the discipline. Finally I try to illustrate both the practical and the ethical limits of avoiding appraisal of alien knowledge systems in a world where the conceptual frameworks and knowledge systems of the researcher and the researched are in constant contact.

2. Avoiding Appraisal

Maria Baghramian (2010) divides the different kinds of relativism that have been influential during the last century into three main groups: conceptual, cultural and postmodern relativism. I shall use this distinction when focusing on the ways in which one particular question is treated in ethnography: How does a researcher encounter different knowledge systems? And, to be precise, how does one treat them in publications? Especially some insights related to conceptual relativism have formed ethnographic research practices into the direction of avoiding appraisal of knowledge systems alien to the researchers' own communities. As postmodern relativism has had an impact on the development of ethnography and pointed cultural research into new directions, it has indeed challenged some earlier practices, but not the one of avoiding appraisal.

Let us understand cultural relativism as the claim that "there can be no such thing as a culturally neutral criterion for adjudicating between conflicting claims arising from different cultural contexts" (Baghramian 2010, 31), and conceptual relativism as the holistic view according to which conceptual frameworks influence thought so strongly that "insofar as it is a question of truth or falsity, one cannot legitimately compare statements made in one [framework] with those made within another" (Mandelbaum 2010, 68). In other words, cultural relativists start the comparison of statements arising from different contexts from a point where it is possible to find them conflicting, whereas conceptual relativists question the possibility of this finding. The first has had a significant role in public discussions about moral and political issues, but the latter has perhaps had a stronger impact on the development of ethnographic research methods and practices. It may be said that whereas some earlier cultural researchers have been (and some contemporary ones still are) cultural relativists and some not, fairly many have been and are – when one looks at their research practices – methodological conceptual relativists.

Wittgenstein and Winch emphasized the need to doubt the applicability of our terminology and norms of rationality when evaluating other knowledge systems. According to them, it is not wise to treat religious practices as mistakes (Wittgenstein 1967) or as unsuccessful scientific hypotheses: "Oracular revelations are not treated as hypotheses and, since their sense derives from the way they are treated in their context, they therefore are not hypotheses." (Winch 1964, 312.) Wittgenstein's remarks were leveled against James Frazer, who did make this kind of comparisons, but much before Wittgenstein wrote his comments, anthropologists had questioned the idea of universal cultural evolution, endorsed by Frazer, as ethnocentric and largely adopted methods where the kind of comparison Wittgenstein criticized is

avoided. Different formulations of conceptual relativism fell into fertile ground amongst ethnographers, and in a moderate form conceptual relativism can be recognized in the ways in which ethnographic research was, and often still is, conducted: Researchers, firstly, accepted the idea that different conceptual frameworks and knowledge systems exist, and secondly, they kept the different systems strictly apart and did not make comparisons between claims made in different systems. The rationale behind this was methodological: Propositions that seem prima facie to be very similar to ones we could make, can, in fact, when made within the unfamiliar conceptual framework, considerably differ from our ways of thinking, and if we presume to be able to understand them well enough right away to make comparisons to our own beliefs, we might not just make a mistake, but in fact hinder our own understanding of the differences in question.[2]

Strong forms of conceptual relativism are problematic, since they can lead to the claim that different conceptual frameworks are incommensurable, which claim turns out to be difficult to defend (Davidson 1974). Ethnographers, who aim precisely at understanding different cultures, and translating between them, cannot accept the idea of full incommensurability and untranslatability between different frameworks. One of the solutions to this problem is to resort to a hermeneutical notion of understanding and interpreting: though the conceptual frameworks of the researcher and that of the researched are different, it is possible to expand the language of the former so as to express the meanings and nuances of the local expressions of the latter – think of Clifford Geertz's "thick descriptions" (Geertz 1973, Risjord 2007). Thus comparisons

[2] On this point even ethnographers who disagree with all forms of relativism, generally agree. For example Dan Sperber, who hardly can be called a relativist, agrees that "resemblances across cultures may well be superficial; failure to understand this leads to poor ethnography" (Sperber 1982, 161).

between statements made in different conceptual frameworks are possible, but only after the slow research process that bridges the gap between the frameworks. However, since the hermeneutic process is often seen as never-ending, and because the research questions of ethnographers often do not necessitate many comparisons, the initial methodological abstinence from critical evaluation can develop into a status quo.

By stressing the significant differences between different conceptual frameworks, and accordingly also different knowledge systems, researchers can at the same time treat their informants' beliefs, ways of argumentation etc. in a respectful manner, and still not take them seriously as propositions that should be accepted, refuted, or compared to the researcher's own claims: for example a Native American myth must not be compared to a scientific hypothesis even if they at first sight might seem to contradict each other. It is the researcher's own academic knowledge system within which theoretical debates happen. One of the most beautifully explicit formulations of this stance comes from Talal Asad:

> Why have I tried to insist in this paper that anyone concerned with translating from other cultures must look for coherence in discourses, and yet devoted so many pages to showing that Gellner's text is largely incoherent? The reason is quite simple: Gellner and I speak the same language, belong to the same academic profession, live in the same society. In taking up a critical stance toward his text I am contesting what he says, not translating it, and the radical difference between these two activities is precisely what I insist on. (Asad 1986, 156.)

I would like to draw attention to two consequences of this differentiation. Firstly, when the beliefs, arguments and ways of producing knowledge claims of the researched are not appraised, they also cannot be adopted and used by the researcher. Of course it is possible to borrow concepts from other conceptual frameworks and add them to the academic arsenal; *mana* and *potlatch* are well-known examples of this. However, academic theoretical discussions do not happen within the informants' conceptual frameworks, nor do ways of knowledge production glide from their knowledge systems to academic argumentation. We do not see for example researchers invoking their age to back up their arguments, even if amongst their informants epistemic authority would be defined by age, nor do we encounter shamanistic research methods. The different knowledge systems are kept quite strictly apart.

Secondly, methodological conceptual relativism is not a practicable stance for researchers who wish to use participatory methods and blur the difference between informants and colleagues. It can indeed be adopted by those who aim at multivocal research: all relevant interest groups are somehow involved in the research process, and get their voice to be heard, but the different stories nevertheless are left clearly apart (e.g. Rountree 2007). However, if participatory research aims further than this, or if the interest groups want more than just to have their story told, too – if they insist on having it accepted as the truth, not just listened to – then methodological conceptual relativism will not do. A methodological conceptual relativist will treat colleagues and informants (or other interest groups) differently.

As noted earlier, it is easy to see how for example Geertz's ideas fit into the description offered here. But avoiding appraisal seems to be a prevalent practice even amongst researchers whose theoretical positions differ from his significantly. For

example, in the more recent constructionistically oriented[3] anthropology and cultural research the focus has been turned towards the researchers' own societies, their conceptual frameworks and knowledge systems. One of the often-criticized concepts is that of culture, especially when used by ethnographers (Wagner 1975), and with it the idea of different knowledge systems being disconnected. Despite this critique, the knowledge systems of the traditional research subjects of ethnography are mostly (though indeed not entirely) left unapprised. The sharp edge of the often social, but sometimes also epistemic critique points to "our", not "their" beliefs and ways of argumentation (Nader 2011). The practice of avoiding appraisal can be and has often been continued within constructionist ethnography.

As mentioned, it is easy to find ethnographic research that incorporates methodological conceptual relativism in its practices. At the same time it seems to be virtually impossible to find ethnographers who would, on the level of their research practices, be consistent epistemic relativists. A consistent epistemic relativist would have no reason not to invoke their age to back up their arguments, if amongst their informants epistemic authority would be defined by age, or to use shamanistic ways of knowledge production in their research. This does not happen. (Koskinen 2011.) Nevertheless, postmodern epistemic relativism has had a strong impact on ethically motivated theoretical debates in anthropology and the neighboring disciplines. It has engendered much discussion on social and cultural inequalities, and it has had an important role in the development of participatory research. This is because of it highlights the relationship between knowledge and power.

[3] By this characterization I refer to Ian Hacking's loose definition. When Marilyn Strathern studies parenthood (2011), or when Regina Bendix studies authenticity (1997), they focus mainly on "our" concepts, tell something about how those concepts have been constructed, and hold that they "need not have existed, or need not be at all as [they are]." (Hacking 1999, 6.)

As Baghramian notes, postmodern relativism is Nietzschean: all knowledge is seen as partial, perspective and tied to power structures, which leads to the conclusion that "we can do little more than insist on the legitimacy of our own perspective and try to impose it on other people." (Baghramian 2010, 45.) Research is seen as inevitably bolstering up one perspective or another, and with it, some power structure. Many cultural researchers inspired by postmodern ideas have concluded that if research is unavoidably political, it should try to unravel existing inequalities and give a voice to the oppressed. This is in dissonance with methodological conceptual relativism, since researchers who actively try to defend marginal ways of thinking and knowing, and empower the communities they are studying, of course take sides and commit themselves much more than a methodological conceptual relativist would find acceptable: knowledge systems are kept less strictly apart, and clearly less emphasis is put on the difficulty of translating. Nevertheless, if appraisal is understood as the act of estimating whether a belief, an argument, or a way of producing knowledge claims is valid or not, postmodern relativism does not encourage researchers to appraise the local thinking they are studying. It does not materially challenge the practice of avoiding appraisal, since the aim is not to appraise beliefs and ways of argumentation, but to empower communities and look for ways in which they could beneficially use their local knowledge. The postmodern researcher quite methodologically *supports* the local knowledge systems, and supporting differs from appraisal.

3. Indigenous Studies: From Research Subjects to Critical Subjects

Avoiding appraisal is possible for researchers as long as the knowledge systems they study can be kept at least somewhat apart from their own knowledge systems. This

was clearly the case in earlier anthropology, where the academic discussion happened far away from the studied people, and it is still the case when the postmodern researchers get to choose what parts of the studied local knowledge they might use in their publications. But the situation is not symmetrical: it has been and continues to be much more difficult for the research subjects to avoid appraisal of the knowledge systems of the researchers. The knowledge produced by researchers is often used in decision-making that affects the lives of the researched, so avoiding appraisal of this knowledge is impracticable. It is not surprising that when the researched have become more acquainted with academic research, some of them have become critical of their role as research subjects. Let us now turn to an extreme example of what happens when research subjects refuse to stay in their role, and want to be treated as simply subjects: the heterogeneous discipline called indigenous studies that claims to bring indigenous knowledge systems into academy.

The notion of *indigenous peoples* has gained significant political weight during the last few decades, much because of the active co-operation of the different activist groups who see themselves as representing the different indigenous peoples around the world. One of their most important agendas has been that of taking control of the ways in which indigenous children and young people are educated. The aim is "the establishment of systems of education which reflect, respect and embrace indigenous cultural values, philosophies and ideologies which have shaped, nurtured and sustained our people for tens of thousands of years." (Seurujärvi-Kari 1996, 171-172.) This includes also higher education and research, and so in different parts of the world there are nowadays colleges and research centers such as the First Nations University of Canada and the Sámi University College, dedicated to research based on indigenous knowledge systems.

The idea of different knowledge systems is generally accepted in indigenous studies, and the prevailing interpretation of it is postmodern[4]: Knowledge is inherently tied to power structures, and researchers who belong to the dominant group and produce knowledge about indigenous people can not easily avoid bolstering up the existing power inequalities. This outlook often involves the Nietzschean idea of understanding a different conceptual framework as a violent act of conceptual appropriation: frameworks are seen as rigid and all-embracing, and understanding means the ruling using their own framework and forcing the ruled to the slots that already exist in it. (Kuokkanen 2006, Tuhiwai Smith 1999, Meretoja 2007.) Despite this it is difficult to find indigenous researchers who would question the applicability of the strongly "western" concept of culture when studying indigenous peoples. Quite the contrary, the notion is used widely and hardly problematised, and it is not difficult even to find "generalizations about the culture as a whole" (Risjord 2007, 416) from indigenous researchers' publications. Given the political force of the concept this is hardly surprising.

The methods used by indigenous researchers often resemble participatory research methods and many research projects are very down-to-earth developmental projects that aspire to engage with the community. Nevertheless, in the theoretical discussions much more controversial ideas have been promoted, such as developing shamanistic research methods (Kuokkanen 2000). The message is altogether clear: indigenous thinking – or "indigenous philosophies" – should be accepted within academy, not "simply as interesting objects of study (claims that some *believe* to be true) but as

[4] There are nevertheless also indigenous researchers who tend to prefer a more Wittgensteinian or hermeneutical approach to the alleged different knowledge systems, and are inclined more towards conceptual than postmodern relativism (Turner 2006, Oskal 2008).

intellectual orientations that map out ways of discovering things about the world" (Garroutte 2003, 10).

The most important aim of indigenous studies is advancing the indigenous identity and self-determination of the indigenous peoples. The main audience is the researcher's own people, so for example a Sámi researcher's work should be directed according to Sámi interests and preferably published in the Sámi language. The openly expressed goal of many of the Sámi researchers is nation-building. (Porsanger 2005, Stordahl 2008, Seurujärvi-Kari 2011.)

Even though indigenous researchers usually understand nation-building as a process of social construction, the building of "imagined communities" (Anderson 1983), the idea of researchers actively building nations is not new. Disciplines such as ethnology and folklore studies have historically had a significant role in the building of some European nation states. Folklorists were notably active in the building of the Finnish nation in the end of the 19th and the beginning of the 20th century. Since then the discipline has gone through an extensive self-critique due to its nationalist history (Anttonen 2005, Wilson 1976). The earlier nationalist research has been deemed dubious in many ways, and the essentialist grounds of the ways in which earlier folklorists represented the Finnish people are seen as especially problematic. Indigenous studies has been criticized similarly: indigenous researchers are said to take *cultures* and *peoples* for granted, and make essentialist assumptions about the studied groups and their local knowledge (Kuper 2003, McGhee 2008).

If an indigenous researcher and a folklorist meet in a conference, it is likely that the latter would like to question some of the theoretical premises of the former. The practical limits of avoiding appraisal become clear: avoiding appraisal of the

indigenous researcher's ideas is in this case not a viable option, even if the folklorist accepts the claim that his ideas stem from an indigenous knowledge system. Either the folklorists expresses her reasons for not agreeing with the indigenous researchers' ideas – thus treating him as a colleague, but taking the risk of apprising a knowledge system which is not her own. Or she stays silent – thus denying the indigenous researcher the status of a colleague who deserves constructive criticism.

4. Participatory Research Needs to Make Room for Debate

Indigenous researchers have achieved something very similar to what participatory research strives for. People belonging to groups that formerly would have been studied by outsider researchers, are now researchers themselves, study their own communities and aspire to base their research methods on their own communities' knowledge systems. Clearly they are not objects of study, but subjects, vis-à-vis other researchers. This has significant consequences for a researcher who accepts the idea of different knowledge systems: indigenous knowledge systems have entered academia, that is, the sphere where critical appraisal of other researchers' ideas is usually encouraged, not avoided. I believe most researchers would agree that "subjecting hypotheses, data, reasoning and background assumptions to criticism from a variety of perspectives" (Longino 2002, 205) is an indispensable part of academic knowledge production, and, accordingly, that it is a researcher's duty to offer criticism to fellow researchers. When indigenous knowledge systems enter academia, the partly ethically motivated practice of avoiding appraisal collides with the ethical obligation of offering criticism to colleagues.

The practice of avoiding appraisal has been continued in participatory research in the name of the ethical imperative of endorsing the knowledge systems of the oppressed. Nevertheless, if the aim of participatory research is to change the relationship between the researcher and the research subject from one between subject and object to one between subject and subject, and to turn informants into collaborators, or effectively even colleagues, the ethical imperative of offering criticism applies to them, too.

When different conceptual frameworks and knowledge systems are in constant contact, research methods and practices that enable the people who see themselves as belonging to different knowledge systems to communicate with each other on a fairly equal footing, subject to subject, are clearly needed. That is, participatory methods and practices are needed. But at the same time, when different conceptual frameworks and knowledge systems are in constant contact, the practice of avoiding appraisal becomes both practically and ethically problematic. In other words, such notions of conceptual frameworks and knowledge systems, as well as such theoretical stances towards them, that do not enable criticism between and across the borders of the different frameworks and systems, are less and less usable in ethnography. They do not lend themselves well to the articulation of the aims of participatory research. Methodological conceptual relativism suffices well for the needs of multivocal research, but not further than that. Postmodern epistemic relativism is inadequate in situations of epistemic conflict. To paraphrase Bernard Williams (1974), when shamanistic ways of producing knowledge claims are no longer only in notional confrontation with academic knowledge production, but have become a real option for researchers, there has to be room for genuine disagreement and debate between the researcher and the shaman.

References

Anderson, Benedict. 1983. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso.

Anttonen, Pertti. 2005. *Tradition Through Modernity: Postmodernism and the Nation-State in Folklore Scholarship*. Helsinki: Finnish Literature Society.

Baghramian, Maria. 2010. "A Brief History of Relativism." In *Relativism, a Contemporary Anthology*, ed. Michael Krausz, 31-50. New York: Columbia University Press.

Bendix, Regina. 1997. *In Search of Authenticity: The Formation of Folklore Studies*. Madison, WI: University of Wisconsin Press.

Carrier, Martin, and Peter Weingart. 2009. "The Politicization of Science: The ESF-ZiF-Bielefeld Conference on Science and Values." *Journal for General Philosophy of Science* 40 (2): 373-378.

Davidson , Donald. 1974: "On the Very Idea of a Conceptual Scheme." *Proceedings and Addresses of the American Philosophical Association* 47 (11973-1974): 5-20.

Finnis, Elizabeth. 2004. "Anthropology and Participatory Research: Ethical Considerations in International Development." *NEXUS* 17(1), article 2.

Garroutte, Eva Marie. 2003. *Real Indians: Identity and the Survival of Native America*. Berkeley: University of California Press.

Geertz, Clifford. 1973. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Hall, Bud L. 2005. "In From the Cold? Reflections on Prticipatory Research From 1970 – 2005." *Convergence* 38(1), 5-24.

Koskinen, Inkeri. 2011. "Seemingly Similar Beliefs: A Case Study on Relativistic Research Practices." *Philosophy of the Social Sciences* 41(1), 84-110.

Kuokkanen, Rauna. 2000. "Towards an 'Indigenous Paradigm' from a Sami Perspective." *The Canadian Journal of Native Studies* 20 (2), 411-436.

Kuokkanen, Rauna. 2006. "The Logic of the Gift: Reclaiming Indigenous Peoples' Philosophies." In *Re-ethnicizing the Minds? Cultural Revival in Contemporary Thought*, ed. Thorsten Botz-Bornstein, and Jürgen Hengelbrock, *Studies in Intercultural Philosophy* 17, 251-271. Amsterdam and New York: Rodopi.

Kuper Adam. 2003. "The Return of the Native." *Current Anthropology* 44 (3): 389-402.

Kurelek, Cathy. 1992. "Anthropological Participatory Research among the Innu of Labrador." *Native Studies Review*, 8(2), 75–97.

Longino, Helen. 2002. *The Fate of Knowledge*. Princeton and Oxford: Princeton University Press.

Mandelbaum, Maurice. 2010. "Subjective, Objective and Conceptual relativisms." In *Relativism, a Contemporary Anthology*, ed. Michael Krausz, 53-79. New York: Columbia University Press.

McGhee, Robert. 2008, "Aboriginalism and the Problems of Indigenous Archaeology." *American Antiquity* 73 (4): 579-597.

Meretoja, Hanna. 2007. "Language – a Bridge or a Barrier? Hermeneutic and Post-Sturcturalist Views on the Possibility of Rational and Ethical Communication." In *Rationality in Global and Local Contexts*, *Reports from the Department of Philosophy*, ed. Jón Ólafsson and Juha Räikkä, 105-120. Turku: University of Turku.

Nader, Laura. 2011. "Ethnography as theory." *Hau, Journal of Ethnographic Theory* 1(1), 211-219.

Oskal, Nils. 2008. "The Question of Methodolgy in Indigenous Research: A Philosophical Exposition." In *Indigenous Peoples: Self-Determination, Knowledge, Indigeneity*, ed. Henry Minde, 331-346. Delft: Eburon.

Park, Peter. 2006. "Knowledge and Participatory Research." In *Handbook of Action Research*, ed. Peter Reason and Hilary Bradbury, 83-93. London, Thousand Oaks, New Delhi: Sage Publications.

Petras, Elizabeth McLean, and Douglas V. Porpora. 1993. "Participatory research: Three models and an analysis." *The American Sociologist* 24 (1): 107-126.

Porsanger, Jelena. 2004. "An Essay About Indigenous Methodology." *Nordlit: Working Papers in Literature* 15, *Special Issue on Northern Minorities*, 105-120. Tromsø: University of Tromsø.

Porsanger, Jelena. 2005. "'Bassejoga čáhci' Gáldut nuortasámiid eamioskkaldaga birra álgoálbmotmetodologiijaid olis." PhD diss., University of Tromsø.

Risjord, Mark 1998. "Relativism and the Possibility of Criticism." *Cogito* 12 (2): 155-160.

Risjord, Mark 2007. "Ethnography and Culture." In *Philosophy of anthropology and sociology*, eds. Stephen P. Turner & Mark W. Risjord, 399-428. Amsterdam and Boston: Elsevier/North-Holland.

Rountree, Kathryn 2007. "Archaeologists and Goddess Feminists at Çatalhöyük: An Experiment in Multivocality." *Journal of Feminist Studies in Religion* 23 (2): 7-26.

Seurujärvi-Kari, Irja. 1996. "Cooperation in the Field of Education and Training among Indigenous Peoples." In *Essays on Indigenous Identity and Rights*, ed. Irja Seurujärvi-Kari & Ulla-Maija Kulonen, 170-178. Helsinki: Helsinki University Press.

Seurujärvi-Kari, Irja. 2011. "'We are no longer prepared to be silent' The making of Sámi indigenous identity in an international context." *Suomen Antropologi* 35 (4): 5-25.

Smith, S. 1997. "Deepening Participatory Action-Research." In *Nurtured by Knowledge: Learning to do Participatory Action Research*, ed. S. Smith, D. G. Willms & N. Johnson, 173-263. New York: Apex Press.

Sperber, Dan. 1982. "Apparently Irrational Beliefs." In *Rationality and Relativism*, ed. Martin Hollis, and Steven Lukes, 149-180. Oxford: Basil Blackwell.

Stordahl Vigdis. 2008. "Nation Building Through Knowledge Building: The Discourse of Sami Higher Education and Research in Norway." In *Indigenous Peoples: Self-Determination, Knowledge, Indigeneity*, ed. Henry Minde, 249-265. Delft: Eburon.

Strathern, Marilyn. 2011."What is a parent?" *Hau, Journal of Ethnographic Theory* 1(1), 245-278.

Tuhiwai Smith, Linda. 1999. *Decolonizing Methodologies: Research and Indigenous Peoples*. London: Zed Books.

Turner, Dale. 2006. *This is not a peace pipe: towards a critical indigenous philosophy*. Toronto: University of Toronto Press.

Wagner, Roy. 1975. *The invention of culture*. Englewood Cliffs, NJ: Prentice-Hall.

Williams, Bernard. 1974. "The Truth in Relativism." *Proceedings of the Aristotelian Society*, New Series, 75:215-228.

Wilson, William A. 1976. *Folklore and nationalism in modern Finland*, Bloomington, IN: Indiana University Press.

Winch, Peter. 1964. "Understanding a Primitive Society." *American Philosophical Quarterly* 1 (4): 307-324.

Wittgenstein, Ludwig. 1967. "Bemerkungen über Frazer's The Golden Bough." *Synthese* 17:233-253.

Wylie, Alison. 2003. "Why Standpoint Matters." In *Science and other Cultures: Issues in Philosophies of Science and Technology*, ed. Sandra Harding, and Robert Figueroa, 26-48. New York and London: Routledge.

Laszlo Kosolosky & Dagmar Provijn

Harvey's bloody motion: Creativity in science

Abstract: In this paper, we show how the discovery of the circulation of the blood by William Harvey (1578-1657) sheds new light on traditional models of creativity in science. In particular, the example illustrates where both the enlightenment and the romantic view on creativity go astray. In the first section, we sketch the two views and present a (non-exhaustive) list of problems for both. In the remainder of the paper, we demonstrate how William Harvey's discovery, as a historical case study of creativity in science, gives firmer ground to these objections.

Our argument goes as follows: First, we show that Harvey is a child of his time as his reasoning is influenced by Aristotle, Galen and the school of Padua (section 2). Second, we indicate how analogies play a considerable role in Harvey's reasoning aside from their usual argumentative value (section 3). Third, Harvey's 'quantitative argument' captures an inherent struggle and reveals a new take on experiments (section 4). Fourth, we elaborate on the dimension of touch in Harvey's use of experiments (section 5). Fifth, vivisection as a research method places Harvey for a dilemma (section 6). Sixth, we engage in the discussion of whether Harvey was an Aristotelian or not, not to solve it, but to argue for his particular historical position (section 7). To conclude (section 8), we spell out the effect of this brief analysis for (A) traditional models of creativity in science and (B) Harvey's historical position.

1. Models of creativity in science

When talking about creativity, one traditionally draws the line between a context of discovery[1], which displays an irrational or 'Eureka' moment, and a context of justification, which exhibits a purely rational dynamic. Over the years several critical remarks arose against Hans Reichenbach's distinction (Reichenbach, 1938).

On the one hand, from 1958 onwards with Norwood Russell Hanson's influential 'Patterns of discovery', the distinction came under attack. Many argued that this distinction needed to be refined by introducing an intermediate step. A third context was supposed to cover both the initial theory formation as well as its preliminary evaluation. Richard Tursman speaks of *"the logic of pursuit and/or of preliminary evaluation of hypotheses"* (Tursman, 1987: 13-14). Ernan McMullin speaks of a *"heuristic appraisal"*, which regards the research-potential of a theory (McMullin, 1976). Larry Laudan describes the intermediate step as *"the context of pursuit"* (Laudan, 1977), and Laurie Anne Whitt as *"theory promise"* or *"theory pursuit"* (Whitt, 1992) (Seselja & Kosolosky, 2012: 1-2). On the other hand, especially since the 1930's Karl Popper and several logical positivists (such as Rudolf Carnap and Carl G. Hempel) took over the distinction and insisted that only matters of justification, and not questions of discovery, obtain its place in philosophical discussion (Nickles, 1980: 1-2).

[1] We understand discovery here rather straightforward as *"a process of thought that leads from (at least partially) observational premises to cognitive conclusions, generally in the form of laws or theories"* (Pera, 1987: 177).

The distinction between context of discovery and context of justification translated itself in two models of creativity in science: Enlightenment and Romantic. In the enlightenment (or classical) model the rational discoverer is endowed with exceptional reasoning skills. Information is sufficient to logically deduce or induce a solution. Behind it lies the idea of the autonomous, individual agent who, in principle, accepts nothing on faith and makes decisions only after an independent application of critical reasoning. The romantic model portrays the discoverer as someone who is sensitive to patterned wholes and a lack of overall fit. At crucial moments in time (s)he experiences a brilliant flash of insight (Eureka) distancing oneself in this manner of common people (Nickles, 1994: 277-278). Standard objections against both models are: (1) they are too individualistic, since they discard the difficulty of locating major historical discoveries. *"The bigger the discovery, the more time it typically takes to work out and articulate the conceptual and instrumental breakthroughs in question, an activity that normally involves many members of the community, including critics"* (Nickles, 1994: 279). (2) They endorse a Whiggish view on science by reading recent developments back into the original observations and concepts. They leave out the role of critical discussion by the larger community over a certain period of time (Nickles, 1994: 279-280). (3) They misrepresent what assigning credit for a discovery entails: *"When scientists assign credit for a discovery, they are doing more than stating an historical fact (that person P discovered that D), they are simultaneously legitimating the corresponding claim and technique, which are usually presented as an extension of an older practice, a continuation of an older tradition"* (Nickles, 1994: 279). (4) There is more not less innovation in science than commonly thought (Nickles, 1994: 280). In the remainder of the paper we investigate in detail how Harvey discovered the circulation of the blood and what brought him to this idea in the first place. This case study serves a dual purpose: on the one hand, it illustrates the theoretically conceived shortcomings of both models of creativity from an historical viewpoint, and, on the other hand, zooming in on these objections allows us to pinpoint the historical Harvey [2].

2. Influenced by…

Harvey, being a child of his time, was influenced by important figures[3], such as Aristotle, Galen, Colombo and Fabricius[4]. According to Aristotle (384-322 BC) the problem of the movement of the heart is the central project for a physician (Pagel, 1944: 145). Harvey, in pursuing this project, took Aristotle's view of the heart as the center of the physiological mechanism (Aird, 2011: 119) and was acquainted with the Aristotelian idea of circular motion as the perfect motion, since there is no motion contrary to it (Pagel, 1944: 145). Galen's (131-207/216) medical doctrine influenced Harvey in at least four ways: (1) Galen introduced the distinction between the venal and arterial system. (2) He endorsed Hippocratic dietetic

[2] We focus on Harvey's main works: 'Prelectionis Anatomiae Universalis' (1616) and 'Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus' (1628) and 'De Generatione Animalium' (1651).
[3] Not only people influenced Harvey, but also tangible inventions in engineering. Boyle claims, in his paper, that a prototype of the sluice gate (the Porte Contarine lock), together with the first pound lock constructed on the Thames, provided Harvey with the decisive model for the function of the venous membranes to obstruct reflux of the blood (Boyle, 2008).
[4] Other examples are Andreas Vesalius (1514-64), Michael Servetus (1511-1553), Andrea Casalpino (1519-1603), Salomon Alberti (1540-1600) and Sanctorius (1561-1636).

and humoral theory in medical practice, based on the 'normality interpretation'[5]. Physicians were thus cautious of learning from dissection (= concerns dead bodies that are not representative for the normal state of the living body) and vivisection (= causes a violent disruption of the normal state of the body). (3) Galen stressed that organs have an attractive force or faculty. In the case of the heart the active process, according to Galen, is the diastole (or expansion) *"[…] during which the heart snatches up or sucks in the inflowing blood like a smith's bellow or sponge."* (Aird, 2011: 121) (4) Galen pinpointed the centrifugal flow of the venal blood, or the flow of the blood from the liver and the heart to outer parts of the body. Another major influence on Harvey was the time spent at the School of Padua, where he interacted with Realdo Colombo (1516-59) and Girolamo Fabricius (1537-1619) (Aird, 2011: 123). Colombo was the first person to portray the pulmonary transit of blood from the right ventricle of the heart to the left. He thus had prior insight in the heart, although his writing was rather ambiguous[6]. Colombo had a dual impact on Harvey: (1) He demonstrated that Galen's work was not devoid of mistakes and (2) he used vivisection as a method to trace these mistakes[7]. Fabricius was the one who prior to Harvey discovered the valves in the veins[8]. This section sketched out some ideas and (minor) discoveries originating in Harvey's (immediate) environment, which, as we will show in the following sections, influenced Harvey in devising the circulation of the blood.

3. Analogies at play

The reason we briefly touch upon Harvey's use of analogies is that they play a crucial role in his reasoning, which surpasses their ordinary argumentative value. We present two examples to illustrate this claim:[9]

I. Analogy as a means to extrapolate

Harvey was able to discern the movement and the action of the heart in fish. But, these observations could not automatically lead to the construction of universals or generalization towards the heart of

[5] A body is considered healthy (or normal) if there is a balance of the four humors, i.e. black bile, yellow bile, phlegm and blood. *"Disease was attributed to an imbalance of humors or a shift in the patterns of flow within the body"* (Aird, 2011: 118).

[6] Colombo confused systole, which is now regarded to be the active movement of the heart (or contraction) with a moment of rest. Diastole was understood as the constriction of the heart (Provijn, under review: 6-7).

[7] We get back to this in section 6.

[8] Both Colombo and Fabricius modified Galen's paradigm, so the *"[…] revival of experimental investigation in the 1500s, while opening the door to progress, did not lead to the downfall of Galen's system of physiology. So persuasive was Galen's theory that these new findings were simply integrated as small modifications into the ancient scheme."* (Aird, 2011: 124) Fabricius, for example, replaced Galen's notion of an 'attractive power', which Galen postulated to keep the blood from falling down into the lower parts of limbs, with a more mechanical explanation. Fabricius thought that ostiola or valves function not as one-way valves (as Harvey later on defended), but only as hindrances to the blood's outward flow. Based on this function, Fabricius argued that the purpose of the valves was to slow the blood's flow, preventing it from collecting too rapidly in the body's extremities (McMullen, 1995: 492).

[9] Other examples are (1) the analogy of the glove (Illustrates the possibility of a passive pulse), and (2) the analogy of the pulmonary transit (The transit of blood may have functioned as an analogue that facilitated to conceive of the transit of blood from the left ventricle throughout the body back to the heart, considering the pulmonary transit as the lesser circulation preceding and contributing to the conception of the full circulation) (Provijn, under review: 13-14).

man, since the structure of the hearts differed considerably. Harvey, guided by the supposition that all hearts have the same function and display analogous processes[10], extrapolates his findings from animal vivisections. To justify this extrapolation, Harvey made use of analogies:

> The same thing is also not difficult of demonstration in those animals that have, as it were, no more than a single ventricle to the heart, such as toads, frogs, serpents, and lizards, which have lungs in a certain sense, as they have a voice. […] Their anatomy plainly shows us that the blood is transferred in them from the veins to the arteries in the same manner as in higher animals, viz., by the action of the heart; the way, in fact, is patent, open, manifest; there is no difficulty, no room for doubt about it; for in them the matter stands *precisely as it would* in man were the septum of his heart perforated or removed, or one ventricle made out of two; and this being the case, I imagine that no one will doubt as to the way by which the blood may pass from the veins into the arteries (Harvey, De Motu Cordis, ch.6: 19, own emphasis).

This extrapolation is unique since Harvey extensively vivisected cold-blooded animals and relied on an unseen number of data to draw conclusions on the possible movement and action of the heart in warm-blooded animals and man, which was a dangerous route to pursue at that time (section 2 and 5).

II. The analogy of the muscle

Harvey draws an analogy between the movement of the heart and the contraction of a muscle:

> […] the motion is plainly of the same nature *as that* of the muscles when they contract in the line of their sinews and fibres; for the muscles, when in action, acquire vigor and tenseness, and from soft become hard, prominent, and thickened: and in the same manner the heart (Harvey, De Motu Cordis, ch.2: 10, own emphasis).

This analogy was never drawn by Colombo and was even opposed by Galen. Galen believed that the heart could not be a muscle since all muscles are held to move with a voluntary motion (Aird, 2011). In this manner Harvey opposed tradition, since the analogy allowed him to suppose the contrary claim that the cavities of the heart must become smaller during systole and that blood is trusted out. He could support this further by the observations he made in fish and other cold-blooded animals.

Analogical reasoning thus serves a larger purpose in Harvey's reasoning: Harvey uses analogical reasoning to draw conclusions on the movement of the heart that certainly occurs in cold-blooded animals (i.e. II) and uses extrapolation to draw the same conclusion for warm-blooded animals (i.e. I). So (I) and (II) combined enabled Harvey to describe the proper movement of the heart and how it related to the propulsive action of the heart, more convincingly than Colombo managed to do before.

4. Quantitative argument[11]

[10] Despite there being morphological differences.

[11] The 'quantitative argument' refers to Harvey's argument against Galen that no matter how much blood is injected into the artery if you multiply this amount by the number of beats per hour of a typical heart it becomes clear that more blood than present in the entire body passes through the heart in a single hour (Lennox, 2006: 18-21).

Harvey experienced drawbacks and difficulties in an attempt to match his new findings on the heart (section 3) with the dominant Galenic cardio-vascular system:

> […] what remains to be said upon the quantity and source of the blood which thus passes is of a character so novel and unheard-of that I not only fear injury to myself from the envy of a few, but I tremble lest I have mankind at large from my enemies, […] And sooth to say, (a) when I surveyed my mass of evidence, whether derived from (b) vivisections, and my various reflections on them, or from the study of the ventricles of the heart and the vessels that enter into and issue from them, the symmetry and size of these conduits, - (d) for nature doing nothing in vain, would never have given them so large a relative size without a purpose, - or (c) from observing the arrangement and intimate structure of the valves in particular, and of the other parts of the heart in general, with many things besides,
>
> I frequently and seriously bethought me, and long revolved in my mind, what might be the quantity of blood which was transmitted, in how short a time its passage might be effected, and the like. But not finding it possible that this could be supplied by the juices of the ingested aliment without the veins on the one hand becoming drained, and the arteries on the other getting ruptured through the excessive charge of blood, unless the blood should somehow find its way from the arteries into the veins, and so return to the right side of the heart, […] (De Motu Cordis, ch.8: 25, own emphasis and introduction of (a), (b), (c) and (d)).

The second half of the quote shows how Harvey, in safeguarding his own theses on the forceful systole and the propulsive action of the heart, had to find a solution to this quantitative problem. His solution: the circulation of the blood, which *"[…] implied that blood was not constantly being consumed in the periphery and replenished by ingested nutrients, but rather that blood was conserved."* (Aird, 2011: 119) Strikingly Harvey argued for his solution by using a thought experiment. He explicitly requests his readers to take an educated guess on the amount of blood that is injected into the arteries. No matter what the answer would be, if we calculate the amount of blood that is to pass through the heart each hour, the exuberant number shows Galen's system to be flawed (Lennox, 2006: 18-19)[12]. This thought experiment must have played a major role in the actual discovery process, for at least two reasons: First, when we judge Harvey's testimony above as trustworthy it illustrates to the historian of science how the thought experiment played a central role in discovery. Second, because Harvey requests an input from the reader by endorsing a 'try this for yourself and see what happens'-mentality (Salter & Wolfe, 2009: 117), we are safe to say that Harvey performed it many times himself (De Mey, 2006: 234-235).

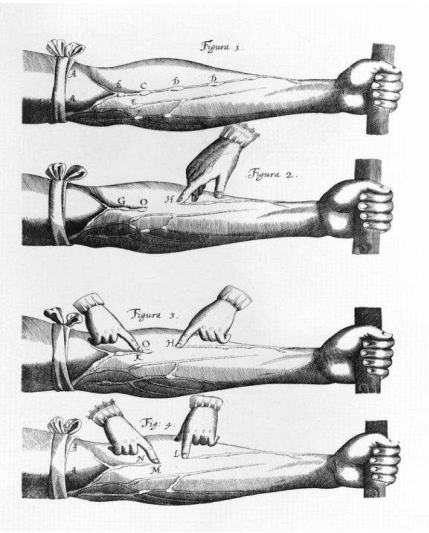
This first half of the quote, however, sheds new light on the empiricist toolbox Harvey considered to be available to him. Harvey uses various (a) equal means of searching for possible solutions, such as (b) experiments, (c) observations and (d) philosophical principles (Salter & Wolfe, 2009: 120). Harvey's view prevents us from pinpointing him as a 'standard' empiricist supporting the distinction between experiments as valuable and observations as inferior knowledge. According to Harvey *"both involve the inspection of nature by sensory perception, both involve active intervention and both require repetition to validate and justify the conclusions they suggest"* (Salter & Wolfe, 2009: 119). Salter & Wolfe argue that labeling Harvey as an empiricist thus calls for a new notion of empiricism, namely medical/embodied empiricism, as opposed to the standard proof-and-validation experimentalism of Royal Society

[12] If we conceive of it this way, one could say Harvey's argument is arithmetic without being quantitative since the exact quantities do not matter.

empiricism (Salter & Wolfe, 2009). Further arguments for this kind of empiricism will be given in the next section, as we present the ligature experiment.

5. Experiment by touch

Harvey took notice of the placing of the valves in the veins and, correspondingly, the flow of the blood in the body by performing the following experiment, which came to be known as the ligature experiment:



Harvey presents this image (Fig. 1) of an arm prepared for bloodletting, i.e., with a ligature tightly bound around it to make the veins swell – with letters marking the position of the valves. In Fig. 2 a finger presses on valve H. The section of the vein below valve H (nearer to the heart) is shown to be empty of blood until the next valve along, valve O. The blood does not flow back through O to fill the vein. In Fig. 3 one finger presses on valve H, stopping the flow towards valve O, while another tries to push the blood from below valve O towards H; yet this section of the vein remains empty – because the valves are stopping the blood from flowing in this direction. However, (Fig. 4) on the opposite side (section M, between valve L and valve N) the vein can fill. Harvey has thus proven that blood flows around the body in one direction only, from the periphery to the heart.

IMAGE Only illustration in William Harvey's De motu cordis, 1628
(Courtesy to the Wellcome Library)

The credibility of touch as an instrument of perception (in line with Harvey's work ethic of doing it yourself and feeling it yourself, see section 4) is, moreover emphatically illustrated by Harvey's inclusion in De Generatione Animalium (1651) of a nobleman Hugh Montgomery as the subject of Harvey's touch. Harvey recounts his live, beating heart which had been exposed by the injuries from a fall when still a child:

> I immediately saw a vast hole in his chest into which I could easily put my first three fingers and my thumb. At the same time I saw just inside the opening, some fleshy, projecting part which was driven backwards and forwards with an alternating movement, and I touched it very cautiously with my hand […] when I had investigated everything carefully enough, it was evident that the old vast ulcer […] was covered over on the inside with a membrane and guarded all round the edges with a hard skin (Harvey, 1651: 250).

The crucial experiment is, again, the experiment of the ligature – but it succeeded because of touch. Only touch could reveal the actual outward and inward flow of blood and the effect of the venous valves in preventing outward venous flow. Salter & Wolfe's conception of empiricism, as set out in section 4, fits

in nicely with Harvey's sense of empiricism, since touch is the crucial characteristic in defining his particular way of experimenting. This kind of empiricism calls for a more 'first person sensitive' perspective, as became evident through Harvey's use of experiments[13].

6. Broadening constraint

Vivisection, as a method of reacting against established theories, served a peculiar role for Harvey. Observing the fast beating of the heart of warm-blooded animals after vivisection did not generate the perspicuous observations needed to draw conclusions on the real active movement of the heart. Harvey was just able to discern the two separate phases, i.e. systole and diastole, but it was impossible to pinpoint one of these as the proper movement. Harvey solved this observation problem by observing cold-blooded animals and dying hearts (section 3 and 4).

From a medical point of view, however, this is both problematic (section 2): (1) The dying heart, plus it being observed during vivisection, could hardly count for the normal situation, and (2) the hearts of cold-blooded animals diverged too much from the ones observed in warm-blooded animals (Provijn, under review: 8). As mentioned in section 4, Harvey tried to bridge this gap through the use of analogies. Vivisection as a method for Harvey could be characterized as, what we call, a 'broadening constraint'. On the one hand, as a natural philosopher it opened up opportunities for him to reject old theories and ground his observations with supplementary power. On the other hand, as a physician it constrained him in using these results and it required supplementary reasoning to convince his fellow physicians, which he found in a thought experiment (section 4), ligature experiment (section 5) and use of analogies (section 3). Understood in this sense, Harvey had the opportunity to be at the interface between being a natural philosopher and being a physician.

7. Conclusion

Our analysis of key factors in Harvey's discovery process sheds light on (A) existing models of creativity in science and (B) Harvey's historical position. Broadly conceived, this paper illustrates how we can understand a historical case in terms of specific characteristics from a philosophy of science perspective, and, vice versa, how a historical case can show us that certain models of scientific discovery and creativity are false, or at least not generalizable.

(A) First, throughout the paper we saw the objections raised against the two models of creativity reemerging:

(1) *Individualistic*: Harvey's ideas were part of a larger community (section 2), governed by critical discussion (section 4 and 5) and a difficulty to pinpoint certain discoveries (e.g. Colombo's part in the discovery of the systole and Fabricius' discovery of the valves in the veins).
(2) *Whiggish*: Since Harvey used the same terminology (e.g. systole, diastole) as his contemporaries, innovation and conceptual change are less straightforward (section 2). His critical interaction

[13] Salter & Wolfe, in understanding empiricism, draw out the following taxonomy: (i) A 'Royal Society', experimentalist empiricism, which may be the context in which an actual 'philosophy of experiment' emerges (i.e. Boyle and Bacon), (ii) A moral/practical empiricism, in which themes such as anti-innatism are in fact not epistemological, that is, not primarily reducible to concerns about the nature of knowledge or of the cognitive states of the knower, but are rather motivated by embedded concerns such as anti-authoritarianism and the desire to articulate a notion of toleration (i.e. Locke and Hume), and (iii) A medically motivated, 'embodied' empiricism (i.e. William Harvey, Pierre Gassendi, Thomas Sydenham) (Salter & Wolfe, 2009).

with Hofmann, Colombo and Fabricius illustrates the presence of critical discussion and assimilation in Harvey's discovery process.

(3) *Assigning credit*: Assigning credit to someone for a discovery entails a legitimation and extension of an older tradition (e.g. Galen, Aristotle, Colombo, Fabricius), combined with new perspectives (e.g. thought experiment, vivisection, role of touch in experiments, analogies).

(4) *More innovation*: Innovation goes in smaller steps (section 2) and experiences restrictions and drawbacks (section 4, 5 and 6), ascribing innovation to more people, since a discoverer is in essence a child of his time.

The example allows us to draw some more general tentative lessons on creativity in science. First, it seems that genuine discoveries are not a product of 'sparks of geniuses' rather than of long, demanding and complex problem-solving processes. Second, the study of creativity should focus on processes and not on products. We have to reconstruct all the elements underlying the process of invention, both historical and formal (i.e. use of analogies) if we aim to address the full extent of the matter. Third, there is nothing magically explanatory about the labels 'enlightenment' and 'romantic', of course both their utility and their potential to mislead should remind us that our ways of describing scientific work, especially innovative research, are tied to larger cultural contexts and are themselves historically conditioned (Nickles, 1994: 308).

(B) Second, we were able to roughly pinpoint the historical Harvey by stressing his interface position in three ways: experimentalist versus Aristotelian (section 2), natural philosopher versus physician (section 6), and embodied empiricist versus experimentalist empiricist (section 4 and 5). Harvey's Aristotelian past, on the one hand, helps him in coming up with possible directions (e.g. heart as central, circular motion), whereas, on the other hand, it also carries constraints (e.g. search for causa finalis). Moreover Harvey was able to combine skills of the natural philosopher and physician in spelling out his circular motion of the blood through, what we would call, an 'experiment/observation grounded thought experiment' (section 6). Last but not least, Harvey is no ordinary empiricist in the sense of Royal Society empiricism, since he places observation and experiment on an equal par (section 4). And so the empirical side of his discovery process consisted mainly of (i) the use of (ligature) experiments and vivisections (section 5 and 6), (ii) the finding of the symmetry and magnitude of the heart ventricles and associated vessels entering and leaving them, (iii) the perceiving of the skillful and careful craftsmanship of the heart valves, fibres and other structural artistry of the heart, (iv) knowing the amount and transmission time of the blood transmitted by the heart, and the fact that the ingested food could not supply this amount without us having the veins (section 4).

It is thus safe to conclude that Harvey represents the struggle between 'the old and the new' and that his intermediary/interface position captures the environment that enabled him to discover the circulation of the blood. This case study, moreover, calls for renewed attention to the study of creativity and discovery in science.

References

Aird, W.C. (2011). Discovery of the cardiovascular system: from Galen to William Harvey, Journal of Thrombosis and Haemostasis, 9: 118-129.

Boyle, M.O. (2008). Harvey in the sluice: from hydraulic engineering to human physiology, History and Technology, 24(1): 1-22.

Bylebyl, J.J. (1982). Boyle and Harvey on the valves in the veins, Bulletin for the History of Medicine, 56: 351-367.

De Mey, T. (2006). Imagination's grip on science, Metaphilosophy, 37(2): 222-239.
French, R. (2004). Harvey William (1578-1657), Oxford Dictionary of National Biography, Oxford University Press.
Hanson, N.R. (1958). Patterns of Discovery: An inquiry into the Conceptual Foundations of Science. Cambridge University Press.
Harvey, W. (1616). 'Prelectionis Anatomiae Universalis'
Harvey, W. (1628). 'Exerciatio anatomica de motu cordis et sanguinis in animalibus'
Harvey, W. (1651). 'Exercitationes de Generatione Animalium'
Laudan, L. (1977). Progress and its Problems: Towards a Theory of Scientific Growth. Routledge & Kegan Paul Ltd, London.
Lennox, J.G. (2006). William Harvey's Experiments and Conceptual Innovation, Medicina & Storia: Rivista di Storia della Medicina e della Sanità: 5-26.
McMullen, E.T. (1995). Anatomy of a physiological discovery: William Harvey and the circulation of the blood, Journal of the Royal Society of Medicine, 88: 491-498.
McMullin, E. (1976). The fertility of theory and the unit of appraisal in science. In Cohen, R.S.; Feyerabend, P.K. & Wartofsky, M.W. (Eds.), Essays in Memory of Imre Lakatos, vol.39 of Boston Studies in the Philosophy of Science, D. Reidel Publishing Company, Dordrecht: 395-432.
Nickles, T. (1980). Scientific Discovery, Logic, and Rationality, Dordrecht: D. Reidel Publishing Company.
Nickles, T. (1994). Enlightenment versus Romantic Models of Creativity in Science – and Beyond, Creativity Research Journal, 7(3&4): 277-314.
Pagel, W. (1944). William Harvey: Some Neglected Aspects of Medical History, Journal of the Warburg and Courtauld Institutes, 7: 144-153.
Pera, M. (1987). The rationality of discovery: Galvanni's animal electricity. In Pitt, J.C. & Pera, M. (Eds.), Rational Changes in Science, 177-201.
Provijn, D. (under review). Bloody Analogical Reasoning.
Reichenbach, H. (1938). Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge, University of Chicago Press.
Salter, A. & Wolfe, C.T. (2009). Empiricism contra Experiment: Harvey, Locke and the Revisionist View of Experimental Philosophy, Bulletin D'Histoire et D'Epistémologie des sciences de la vie, 16(2): 113-140.
Seselja, D. & Kosolosky, L. (2012). Rationality of Scientific Reasoning in the Context of Pursuit: Drawing appropriate distinctions, Philosophica, 88
Tursman, R. (1987). Peirce's theory of scientific discovery. Indiana University Press. Bloomington and Indianapolis.
Wear, A. (1983). William Harvey and the 'way of the anatomists', History of Science, 21: 223-249.
Whitt, L.A. (1992). Indices of theory promise, Philosophy of Science, 59: 612-634.

1

Understanding non-modular functionality – lessons from genetic algorithms

Jaakko Kuorikoski and Samuli Pöyhönen[1]

University of Helsinki

Abstract

Evolution is often characterized as a tinkerer that creates efficient but messy solutions to problems. We analyze the nature of the problems that arise when we try to explain and understand cognitive phenomena created by this haphazard design process. We present a theory of explanation and understanding and apply it to a case problem – solutions generated by genetic algorithms. By analyzing the nature of solutions that genetic algorithms present to computational problems, we show that the reason for why evolutionary designs are often hard to understand is that they exhibit non-modular functionality, and that breaches of modularity wreak havoc on our strategies of causal and constitutive explanation.

1 Introduction

The once dominant classical paradigm of cognitive science has been under attack for several decades. Connectionism, cognitive neuroscience, dynamical systems theory, and new robotics have all questioned whether the classical AI approach to cognition can credibly describe biologically evolved cognitive systems such as human minds. Whereas classical AI tends to approach computational problems with functional decompositions inspired directly by the programmer's intuitions about possible efficient subroutines, the alternative research programs often emphasize that biological evolution is more likely to produce far more complex and messy designs.

In our paper we analyze the nature of the problem that these messy solutions raise to the understanding of cognitive phenomena. In general, the problem of understanding non-intuitive designs produced by natural selection is well-known in philosophy of psychology (e.g., Clark 1997, Ch. 5), philosophy of biology (Wimsatt 2007), and now even in popular psychology (Marcus 2008), but the problem has proven to be difficult to articulate without a clear idea of what exactly it is that

[1] The authors are listed in an alphabetical order.

evolutionary tinkering is supposed to hinder. The main challenge for understanding is often framed and explained by pointing to the path-dependent nature and the resulting unfamiliarity of the evolved design (Jacob 1977). We argue that this is not the whole story. We hope that providing an explicit theory of explanation and understanding will move us beyond intuitions towards a more systematic analysis and, ultimately, concrete solutions. We also combine our theory of explanatory understanding with a computational application of evolutionary design: problem-solutions generated by genetic algorithms. By analyzing the nature of solutions that genetic algorithms offer to computational problems, we suggest that an important reason for why evolutionary designs are often hard to understand is that they can exhibit non-modular functionality, and that breaches of modularity wreak havoc on our strategies of causal and constitutive explanation.

2 Explanation and understanding

The ultimate goal of cognitive neuroscience is to provide mechanistic understanding of system-level properties of the cognitive system in terms of the properties of its parts and their organization. Probably the most developed account of general strategies for reaching such mechanistic understanding is William Bechtel's and Robert Richardson's (2010) study of the heuristics of decomposition and localization (DL). The DL procedure goes roughly as follows. First, the different phenomena that the system of interest exhibits are differentiated. Then the phenomenon of interest is functionally decomposed, i.e., analyzed into a set of possible component operations that would be sufficient to produce the phenomenon. One can think of this step as a formulation of a preliminary set of simple functions that taken together would constitute the more complex input-output relation (the system-level phenomenon). The system is also structurally decomposed into a set of component parts. The final step is to try to localize the component operations by mapping the operations onto appropriate structural component parts. The idea is thus to first come up with a set of more basic properties or behaviors which could, taken together, possibly result in the explanandum behavior, and then try to find out whether the system is in fact made of such entities that can perform the required tasks. If this cannot be done, the fault may lie with the functional and structural decompositions or with the very identification of the phenomenon, and these may then have to be rethought. The identification and decomposition procedures will in the beginning be guided by earlier theories and common sense, but empirical evidence can always suggest that a thorough reworking of the basic ontology and the form of the possible explananda may be in order.

According to Bechtel and Richardson, decomposability is a regulative ideal in such model construction because complex systems are psychologically unmanageable for humans.

Decomposition allows the explanatory task to be divided into parts that are manageable for cognitively limited beings, thereby rendering the system intelligible (Bechtel and Richardson 2010). The idea comes originally from Herbert Simon (1962), who claimed that the property of near-decomposability is a necessary condition of understandability to any finite cognitive agent. Near-decomposability means that the system can be decomposed into parts in such a way that the intrinsic causal properties of the parts are more important for the behavior of the system than the relational causal properties of the components that are constituted also by their environment and interaction. Near-decomposable systems are thus hierarchical in the sense that the complex whole can be conceived of as made from a limited set of simpler parts and interactions. Hierarchical systems are manageable for cognitively limited beings because their 'complete description' includes irrelevant elements describing similar recurring parts and non-important interactions. The removal of such descriptions does not hamper our understanding of the system and thus eases cognitive load.

Although there are a number of arguments that conclusively show that such informational economy by itself is not constitutive of understanding[2], we agree with Simon in that a property closely related to near-decomposability, namely modularity, is a necessary condition for understanding. As a conceptual starting point for our argument, we follow Petri Ylikoski and Jaakko Kuorikoski in conceiving understanding not as a special mental state or act, but as a regulative label attributed according to manifest abilities in action and correctness of reasoning. Understanding is a public, behavioral concept. Cognitive processes (comprehension) taking place in the privacy of individual minds are a causal prerequisite for possible fulfillment of these criteria, but the processes themselves are not the facts in virtue of which somebody understands or not. They are not the criteria of understanding in the sense that we would have to know them in order to say whether somebody really understands something. (Ylikoski 2009; Ylikoski and Kuorikoski 2010)

We take the primary criterion of understanding to be inferential performance: whether someone understands a concept is evaluated according to whether he or she can make the right inferential connections to other concepts. Likewise, whether someone understands a phenomenon is assessed based on whether he or she can make correct inferences related to it. This view can be further

[2] First, nobody has actually succeeded in giving a positive argument for equating understanding with increased informational economy (Barnes 1992). Second, successful classification schemes compress information by facilitating inferences to properties probably possessed by individuals on the basis of belonging to a certain known class. However, classification schemes by themselves are usually taken to be merely descriptive and not explanatory. The same general point can be drawn from standard statistical procedures, which by themselves only summarize the data, but do not explain it. (Woodward 2003, 362-364.)

developed by linking it to James Woodward's account of scientific explanation in the following way: Woodward's theory of explanation tells us more specifically what kinds of inferences are constitutive of specifically explanatory understanding. According to Woodward (2003), explanation consists in exhibiting functional dependency relations between variables. Knowledge of explanatory relationships facilitates understanding by implying answers to what-if-things-had-been-different questions concerning the consequences of counterfactual or hypothetical changes in the values of the explanans variable. Whether someone understands a phenomenon is evaluated according to whether he or she can make inferences not only about its actual state, but also about possible states of the phenomenon or system in question. In the case of causal explanations, these explanatory dependencies concern the effects of interventions and knowledge of causal dependencies thus enables the possessor of this knowledge to act and possibly manipulate the object of explanation. These answers are the basis of the inferential performance constitutive of understanding.

The limits of inferential performance depend causally on contingencies related to the reasoning processes of the agents whose understanding is being evaluated. Thus the limits of understanding are dependent on the cognitive make-up of agents and can certainly be investigated psychologically. For example, if the space limit of our working memory is indeed roughly seven items, then this constitutes an upper boundary for the complexity of our inferences and, consequently, for our understanding.

In order for answers to what-if questions to be well defined, the dependencies grounding the answers have to possess some form and degree of independence such that a local change in an aspect of the phenomenon under study cannot ramify uncontrollably or intractably. If local modifications in a part of a system disrupt other parts (dependencies) in a way that is not explicitly specified (endogenized) in the (internal or external) representation of the system according to which the what-if inferences are made, the consequences of these changes are impossible to predict and counterfactual assertions impossible to evaluate. Things participating in the dependency relations also have to be somewhat localized (physically and/or conceptually) in order for the contemplated changes to be well defined in the first place. (Woodward 2003, 333.) Therefore a necessary condition for a representation to provide understanding of a phenomenon is that the modularity in the representation matches the modularity in the phenomenon.

Let us first discuss the case of causal understanding. If an intervention on a causal system actually changes the system in a way that is not represented in the model of the system, the model as it stands does not give correct answers to what-if-things-had-been-different questions concerning the state of the system after the intervention. If we intervened on a causal input corresponding to

variable X_i in a model and the intervention, no matter how surgical, also changed the dependencies within the system or values of other variables themselves affecting variables causally downstream of X_i, the model would give incorrect predictions about the consequences of the intervention. Hence, the model would not provide correct causal understanding of the workings of the system and the causal role of the variable in it. If the system cannot be correctly modeled on any level of description or decomposition so that it is modular in such a way – if the system itself is not causally modular – no what-if-things-had-been-different questions concerning interventions in the system can be answered and there is no causal understanding of the system to be had. If the system is in fact such that every local change brings about intractable changes elsewhere in the system to such an extent that there can be no representation that would enable a cognitively finite being to track these changes and make correct inferences about their consequences, then the system is beyond the limits of understanding.

The problem of understanding causally non-modular systems has received some attention in the philosophy of science literature (e.g., Bechtel and Richardson 2010, Ch. 9). However, according to the schema of Bechtel and Richardson, before we can even start thinking about acquiring causal-mechanical understanding of the system realizing the complex behavior to be understood, we need to formulate hypotheses about the possible functional decompositions of the behavior (see also Cummins 1983). For example, what kind of simpler subtasks could possibly produce complex cognitive capacities such as language production and comprehension, long-term memory, and three-dimensional vision? Importantly, these hypotheses are separate, though not independent, from hypotheses concerning the implementation of the capacity. Although the understanding offered by the functional decomposition is not strictly speaking causal – component operations do not cause the whole behavior because they are constitutive parts of it[3] – the modularity constraint on understandability still applies in the following way. We can only understand the complex behavior by having knowledge of the component operations if we can make reliable what-if inferences concerning the possible consequences of changes in the component operations for the properties of the more complex explanandum capacity. We provisionally understand working memory if we can infer from possible changes in its hypothesized component operations (such as differences in the postulated phonological loop or episodic buffer) to changes in the properties of the capacity. These inferences are only possible if the functional decomposition itself is suitably modular, i.e., the consequences of "local" changes in component operations do not ramify in an intractable way

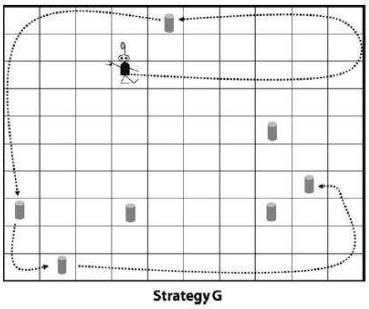
[3] Although we fully agree with Piccinini and Craver (2011) in that insofar as functional decompositions are explanatory, they are to be thought of as mechanism sketches and that the functional hypotheses are not independent of the question of mechanistic implementation.

making the behavior of the whole completely holistic. We now argue that genetic algorithms demonstrate that design by selection can lead to such non-modular complex behavior.

3 Genetic algorithms

Since the 1960s, there have been attempts to apply insights from evolutionary thinking to computer programming. Here we discuss one genre of evolutionary programming: genetic algorithms (cf. Holland 1975; Goldberg 1989; Mitchell 1996). In a nutshell, the idea of the genetic-algorithms approach is to "breed" randomly generated solutions to computational problems. This is done by mimicking the evolutionary mechanisms of inheritance, mutation, selection and crossover in a computer simulation. Although genetic algorithms (henceforth GAs) are not the only strand of evolutionary programming, they serve our purpose well because their basic principles are easy to understand and they are the most well-known kind of evolutionary programming outside computer science (Clark 1997, 2001; Mitchell 2009).

From the point of view of AI, genetic algorithms are a form of non-exhaustive but massively parallel search in the search space of a problem. They can be used for a number of different purposes: for evolving behavioral strategies for simulated agents, for finding weights for a connectionist network, or for evolving cellular automata to perform computations. We illustrate the nature of GAs by presenting a simple example from Melanie Mitchell (2009, Ch. 9). Mitchell's original simulation showed how GAs can be used to evolve a controlling program for a simulated robot picking up soda cans in a 10x10 grid. Robby the robot can only see squares that are adjacent to its location (center, North, South, East, West), and each turn it can either move one step to a particular direction, move at random, try to pick up a can, or do nothing. Each simulation run lasts for a predetermined amount of time steps (originally 200), and Robby's task is to pick up as many randomly situated soda cans as possible.



Strategy G

Genome G:

```
25435515325623525105635546115133615415103415611055015005203025625613225235032511205233305405523125505133615415066526415026650601226445360563152025643105435463240435033415325025325135235204515013015621343625235322313505126051335620152451434343232
```

Figure 1. (taken from Mitchell 2009, 137). Each "locus" in the genome G corresponds to one of the possible immediate environmental states of Robby and each digit (the allele) to a move in that situation (e.g. '0' → 'move north', '5' → 'pick up').

Initially a random population of software individuals is generated, each with a "genome" consisting of 243 random numbers. Each locus in the genome guides Robby's behavior in a particular situation (Fig 1). The fitness score of each candidate program in the population is calculated by running several simulation trials: crudely, the more cans the robot is able to pick up by average, the higher its fitness. Programs with the highest fitness scores are then used to form the next generation of programs: they are paired randomly, and the genomes of the two parents are crossed over at a randomly chosen point to create the genomes of new individuals. Finally, for each descendant, there is a small probability (.05) that a mutation occurs in its chromosome. As a result, the new generation is based on the most successful variants among the previous generation and the process loops back to the fitness-calculation phase. Thus the GA continues searching for efficient solutions to the problem by investigating the surrounding areas in the search space.

After a few hundred generations, the evolved strategies start to achieve impressing results in the simulated task. As we replicated Mitchell's simulation, we observed that after the 800th generation, the best strategies among evolved Robbys started to have higher fitness scores than a simple "rational" solution programmed by a human designer (ultimately 480 vs. 420 points). However, although solutions found with GAs are efficient, their behavior is often hard to understand. The ingenious heuristics that the programs employ cannot be deciphered by simply looking at individual genes or sets of genes. Instead, looking holistically at the broad phenotypic behavior of the robot is necessary. A nice illustration of this impenetrability of such evolved solutions is the fact that in some cases when a highly evolved Robby is in the same square with a can, it decides not to pick it up, but rather chooses to move away from the square. While this behavior seems prima facie irrational, looking at the total behavioral profile of the robot uncovers a cunning strategy: Robby uses cans as markers to remember that there are cans on its side and explores the adjacent squares for extra cans before picking up the marker can. Thus by not treating cans only as targets but also as navigational tools, Robby uses its environment to extend its severely limited visual capacities and to compensate for its total lack of memory.

Moreover, by examining the behavior of a 1500th generation Robby that has the highest fitness score in its population, it can be seen that the marker strategy manifests in slightly different ways in different environmental situations. It is therefore not a discrete adaptation, but rather a collection of independently evolved sub-strategies. Furthermore, the marker strategy appears to tightly intertwine with other environment-employing "hacks" that the sophisticated Robby uses: when there is already a lot of empty space on the grid, Robby employs a "vacuum-cleaner" movement strategy. It follows the walls of the board, departing toward the center when it detects a can, employs the marker strategy if possible, and immediately after cleaning up its local environment, returns directly to the south wall to continue its round around the board.

Such kluges are common to designs created by GAs. Like biological evolution, GAs can come up with solutions that a human designer would not usually think of. These solutions often offload parts of problem solving to the environment, and thus rely on a tight coupling between the system and its environment. And as pointed out by Clark (1997, 2001), recurrent circuitry and complex feedback loops between different levels of processing often feature in systems designed by GAs. Such designs are often difficult to understand. We claim that such difficulties in understanding are often created by the lack of modularity in the functional decomposition of the behavior. This point can be illustrated by looking again at the genome of our most successful Robby (genome G in Fig 1). Robby is leaving cans as markers only in specific situations and only the totality of this selective marking strategy, together with navigational strategies utilizing cans and walls, constitutes the effectiveness of the search procedure. Looking at isolated genes in Robby's genome only reveals trivially modular elements corresponding to elementary subtasks in Robby's behavior: one gene corresponds to an elementary move in a specific environmental situation. But we cannot make inferences from local hypothetical changes in these elemental behaviors to consequent effects on fitness. The connection between any single elementary behavioral rule and the strategy is simply too complex and context dependent. A change in a single rule (in situation B and a can present, whether to pick or not to pick the can up) has consequences for the effectiveness of the other elementary behavioral rules constituting the navigational strategy. Explanatorily relevant inferences would require an extra "level" of modular sub-operations between the individual movements and the strategy as a whole. The marker and vacuum-cleaner strategies mentioned above are examples of such middle-level sub-operations, but they are by themselves insufficient to yield understanding of the whole behavior of our most successful Robby, since the effectiveness of leaving a can is a result of the evolved match between the specific situations in which Robby leaves a can and the rest of the navigation behavior. And genetic algorithms do not, in general, produce such easily discernible designs. Rather, only by

simultaneously looking at constellations of different genes, and eventually the whole genome, the interesting heuristics in the system's behavior can be revealed - if at all.

To recapitulate, our example exhibits several distinct (yet related) challenges to understanding:

1. The discernible middle-level strategies (marker, vacuum cleaner) do not have a dedicated structural basis. Instead, the nature of the design process leaves all atomic structural elements (the 243 DNA elements) open for exploitation by all capacities serving the main goal. In consequence, the system is not structurally or behaviorally nearly-decomposable, but instead has "a flat hierarchy." Strategies are implemented in highly distributed structures, and as pointed out in section 2, this raises a challenge for human cognitive capacities.
2. Challenge 1 above means that the interactions between subtasks tend to be strong: a change in one subtask constituting a part of the marker-behavior affects also the functioning of the vacuum-cleaner navigation. In general the middle-level strategies can only be discerned and defined in a very abstract way and the interaction-effect in their contribution to the overall fitness is so large as to make any inferences about the consequence of partial changes in one strategy next to impossible.
3. The way in which operations contribute to the fitness of the individual is highly context-dependent and depends on the properties of the environment as well as the DNA of the agent. For instance, merely detecting the existence of the marker strategy requires that there are suitable clusters of cans in the environment. Moreover, even small modifications to the environment can lead to drastic changes in the performance of a strategy. For instance, adding only a few randomly placed extra walls on the grid radically collapses the average score of the successful Robby described above.

Extrapolating from this very simple case, GAs may yield functional decompositions of the problem that do not follow a tidy hierarchical decomposition into modular subtasks, whose individual contributions would be easy to understand (i.e., we could infer how a change in a sub-routine would affect the behavior of the mother-task). Instead, feedback, many tasks using same subtasks as resources, and environment couplings lead to holistic design where almost "everything is relevant for everything." The evolved functional architecture is flat in that there are few discernible levels of order between the elementary operations and the complex whole. The counter-intuitiveness of such flat architectures is apparent in the deep mistrust faced by connectionist suggestions for non-

hierarchical design of cognitive capacities (see e.g., Rumelhart and McClelland 1986 vs. Pinker and Prince 1988).

Furthermore, GAs underscore the path dependence of evolutionary problem solving. For sufficiently complex computational problems there are often several local maxima in the fitness landscape of the problem, and the population can converge to different maxima in different runs of the simulation. The functional decomposition that a human designer comes up with is just one possible solution among several others. Perhaps our biological evolution actually ended up with a radically different one.

4 Lessons for the study of mind

Genetic algorithms seem to demonstrate that evolution can in principle lead to non-modular functionality. This imposes a limit on our ability to understand such behavior: if we cannot trace the consequences of changes in the sub-operations, we cannot answer what-if questions concerning the complex behavior. Such behavior also constitutes a thorny problem for mechanistic understanding of the implementation of the said behavioral capacities, since the DL heuristic cannot even get off the ground. We can now ask two questions: should we expect to find such non-modular functionality in nature, especially in human cognition, and if so, what attitude should we adopt with respect to this problem. Should the aim of causal-mechanistic understanding of the brain be given up and replaced with a program of instrumentally interpreted dynamical models and modeling the dynamics of the mind with a few macro-variables?

There are important disanalogies between GAs and biological evolution. (1) in GAs, there usually is no genotype–phenotype distinction. In biological evolution, however, genes do not directly cause properties of the phenotype, but rather participate in guiding ontogenesis. There have been suggestions that ontogenesis itself favors modular design. GAs may also seem a problematic platform for exploring the possibilities of DL heuristics, since the lowest level of functional organization and the level of implementation are the same (i.e., the genome). However, we see no reasons why this would affect our argument. Moreover, the argument developed here is about selection in general, and failures of functional modularity may in principle also arise in the course of development – at least if the idea of neuronal group selection or "neural Darwinism" is taken seriously.

(2) Most studies on genetic algorithms are carried out by using a single fixed goal or a fixed task type. In the Robby example, although the distribution of the cans was generated at random, the task itself remained essentially the same from generation to generation. However, Nadav Kashtan and Uri Alon (2005, see also Kashtan et al. 2007) have demonstrated that when the goals themselves are composed of modularly varying sub-goals, evolution produces modular functionality. It is easy to see why this is the case. If the tasks to which the system has to adapt to remains the same, the selection environment is stable and the peaks in the fitness landscape are immovable, then selection favors strategies which offload problem solving to that particular environment as much as possible. But if the task itself is composed of changing subtasks, it makes sense to design the adaptive response in such a way that a particular sub-operation can locally adapt to a local change in a subtask without altering the totality of the otherwise well functioning behavior.

It seems likely that cognition has evolved in such a modularly changing selection environment, but the extent to which we should expect to find modular functionality in human cognition is hard to estimate and is most probably a purely empirical matter. Moreover, as a response to Simon's (1962) Tempus and Hora argument, it has been argued that componential specialization in complex systems is a force that works against the development of strictly modular structures (e.g., Levins 1973, Wimsatt 2007, 186–192). Nonetheless, these arguments as such give us no reason to believe that the produced functional decomposition should respect any intuitive constraints, such as those derived from introspection on our thought processes or the way in which we would program a strategy to tackle similar cognitive challenges.

Genetic algorithms demonstrate that evolution can create designs which are in principle beyond the understanding of unaided cognitive beings such as us. Yet there is nothing mysterious in such designs. Simon pondered whether the relative abundance of hierarchical nearly decomposable complexity was due to our selective attention to precisely such systems, but we believe this to be a somewhat hasty conjecture. We have no trouble finding and delineating systems, such as Robby or possibly ourselves, with behaviors which are functionally non-decomposable and constituted by a flat architecture. However, there certainly might be a psychological bias that makes us see hierarchical design also where there is none. One way of coping with this impasse is to realize that there are no fundamental reasons to limit the relevant understanding epistemic agent to be an unaided human. Although only a human agent can experience a sense of understanding, this feeling should not be confused with understanding itself. Therefore brute computational approaches can

produce understanding as long as the understanding subject, the cognitive unit whose inferential abilities are to be evaluated, is conceived as the human-computer pair.

References

Barnes, Eric. 1992. Explanatory Unification and Scientific Understanding. PSA 1992, 3–12.

Bechtel, William and Robert C. Richardson. 2010. Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research, The MIT Press.

Clark, Andy. 1997. Being There: Putting Brain, Body and World Together Again. The MIT Press.

Clark, Andy. 2001. Mindware: An Introduction to the Philosophy of Cognitive Science. Oxford University Press.

Cummins, Robert. (1983). The Nature of Psychological Explanation. Cambridge: MIT Press.

Goldberg, David E. 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison Wesley.

Holland, J. 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press.

Jacob, Francois. 1977. Evolution and Tinkering. Science 196 (4295): 1161–1166.

Kashtan, Nadav and Uri Alon. 2005. Sponatenous evolution of modularity and network motifs. PNAS 102 (39), 13773–13778.

Kashtan, Nadav, Elad Noor and Uri Alon. 2007. Varying environments can speed up evolution. PNAS 104 (34), 13711–13716.

Levins, Richard. 1973. The Limits of Complexity. Pattee, H. (ed.) Hierarchy Theory: The Challenge of Complex Systems. London: Braziller: 73–88.

Marcus, Gary. 2008. Kluge: The Haphazard Construction of the Human Mind. Boston and New York: Houghton Mifflin.

Mitchell, Melanie. 1996. An Introduction to Genetic Algorithms. Cambridge: MIT Press.

Mitchell, Melanie. 2009. Complexity. A guided tour. Oxford: Oxford University Press.

Piccinini, Gualtiero and Carl Craver. 2011. Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches. Synthese 183 (3):283–311.

Pinker, S. and Prince, A. 1988. On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. Cognition 23: 73–193.

Rumelhart, D. and McClelland J. 1986. On Learning the Past Tenses of English Verbs. in McClelland and Rumelhart et al. 1986. Parallel Distributed Processing, vol. I, Cambridge, Mass.: MIT Press: 216–271.

Simon, Herbert. 1962. The Architecture of Complexity. Proceedings of the American Philosophical Society 106. 476–482.

Wimsatt, William. 2007. Re-Engineering Philosophy for Limited Beings. Cambridge MA: Harvard UP.

Woodward, James 2003. Making Things Happen. Oxford University Press.

Ylikoski, Petri. 2009.The Illusion of Depth of Understanding in Science. in Scientific Understanding: Philosophical Perspectives (edited by H. De Regt, S. Leonelli & K. Eigner), Pittsburgh University Press: 100–119.

Ylikoski, Petri and Jaakko Kuorikoski. 2010. Dissecting Explanatory Power. Philosophical Studies 148, 201–219.

What scientists know is not a function of what scientists know

P.D. Magnus*

June 13, 2012

This has been paper accepted for presentation at the Philosophy of Science Meeting in San Diego (November 2012) with the final version to be published in a supplementary volume of *Philosopy of Science*.

Abstract

There are two senses of 'what scientists know': An individual sense in which scientists report their own opinions, and a collective sense in which one reports the state of the discipline. The latter is what is of interest for the purpose of policy and planning. Yet an expert, although she can report the former directly (her opinion on some question), can only report her considered opinion of the latter (the community opinion on the question). Formal judgement aggregation functions offer more rigorous frameworks for assessing the community opinion. They take the individual judgements of experts as inputs and yield a collective judgement as an output. This paper argues that scientific opinion is not effectively captured by a function of this kind. In order to yield consistent results, the function must take into account the inferential relationships between different judgements. Yet the inferential relationships are themselves matters to be judged by experts involving risks which must be weighed, and the significance of the risk depends on value judgements.

In one sense, 'what scientists know' just means the claims which are the determination of our best science. Yet *science* is a collective enterprise; there

*Thanks to John Milanese, Heather Douglas, and Jon Mandle for comments on various parts of this project.

are many scientists who have individual and disparate beliefs. So 'what scientists know', in another sense, means the omnibus comprised of the epistemic state of scientist #1, the epistemic state of scientist #2, and so on for the rest of the community. The phrase is ambiguous between a collective and an individual meaning.

If we consult a scientific expert, either because we want to plan policy or just because we are curious, we are typically interested in the collective sense. We want to know what our best present science has to say about the matter. And the expert we consult can differentiate the two senses, too. She can relate what she as a particular scientists knows (what she herself thinks, where here sympathies lie in controversies, and so on), but she can also take a step back from those commitments to give her sense of what the community consensus or dominant opinion is on the same matters. If it is simply curiosity that has led us to consult an expert, this may be enough. When policy hangs on the judgement, however, we want more than just one expert's report on the state of the entire field.

This distinction between their personal commitments and the state of the field in their discipline is one that any scholar can make. If you think (as tradition has it) that only individuals can have *beliefs* in a strict sense, then take the expression 'opinion of the scientific community' as a *façon de parler*. If you think (as Lynn Hankinson Nelson does [10]) that the community rather than the individual *knows* in a strict sense, then suitably reinterpret 'what an individual knows' in terms of belief. The distinction I have in mind is neutral with respect to the metaphysics of social epistemology. The question is simply how we could use consultation with individuals to generate a composite, collective judgement.

Formal *judgement aggregation* offers rigorous frameworks which seem to provide what we want. In the abstract, it defines a function that takes individual scientists' judgements as inputs and yields collective judgement as an output. This assumes that the collective judgement of the scientific community depends on the separate individual judgments of the scientists — i.e., that *what scientists know* in the collective sense is a function of *what scientists know* in the individual sense.

Taking a recent proposal by Hartmann et al. [6][7] as an exemplar, I argue that judgement aggregation does a poor job of representing *what scientists know* in the collective sense. I survey several difficulties. The deepest stems from the fact that judgements of fact necessarily involve (perhaps implicit) value judgements. Where values and risks might be contentious, this entails

that individual judgements cannot merely be inputs to a function. Judgement aggregation is not enough.

1 The majority and premise-majority rules

As a judgement aggregation procedure, one might naïvely survey scientists about factual matters and take any answer given by the majority of scientists to reflect the state of science. Of course, scientists would agree about a great many things that are simply not within their purview. Physicists would say that Sacramento is the capital of California, but that does not make it part of physics. So the survey should be confined to matters that are properly *scientific*. The survey must also include only legitimate scientists and exclude ignorant rabble. These restrictions are somewhat slippery, but let's accept them.

The naïve procedure is a simple function from individual judgements to an aggregate judgement: Return the judgement endorsed by a majority of the judges. Call this the *majority* rule.

The *majority* rule has the nice features that it treats every judge equally and that it does not bias the conclusion toward one judgement or another. Yet it suffers from what's called the *discursive dilemma*: It can lead to inconsistent collective judgements, even if all the judges considered individually have consistent beliefs. In the following schematic example, there are three judges: Alice, Bob, and Charles. Each has the consistent beliefs on the matters P, Q, and $(P\&Q)$ indicated in the table below. The *majority* rule yields the inconsistent combination of affirming P and Q but denying $(P\&Q)$.

	P	Q	$(P\&Q)$
Alice	T	F	F
Bob	F	T	F
Charles	T	T	T
majority	T	T	F

The nice features of *majority* rule seem like desiderata for a judgement aggregation rule, but avoiding the discursive dilemma is another such desideratum. A good deal of ink has been spilled specifying precisely the desiderata and proving that they are together inconsistent. However, even where it can be proven that a set of desiderata cannot be satisfied in *all* cases, they may

still be jointly satisfied in some instances. The *majority* rule can lead to contradiction, it does not do so in every case. As a practical matter, we might begin by trying out a simple rule (like *majority*) and add sophistication only if the actual community has judgements like those in schematic example.[1] Even so, more sophisticated rules would be needed for corner cases.

Stephan Hartmann, Gabriella Pigozzi, and Jan Sprenger [6][7] develop a judgement aggregation rule specifically to escape the discursive dilemma. Their procedure involves polling judges only regarding matters of independent evidence. For matters which are consequences of the evidence, the procedure derives consequences from the aggregated judgements. In the simple case given in the table above, for example, the procedure would affirm P and Q (because each is affirmed by a majority) and also $P\&Q$ (because it is a consequence of P and Q). Call this the *premise-majority* rule. When it can be applied, *premise-majority* generates a consistent set of judgements.

There are several difficulties with *premise-majority*, as a way of aggregating expert scientific opinion.[2]

First, *premise-majority* inevitably produces some determinate answer. As Brams et al. [3] show, it is possible for a combination of separate elections to result in an overall outcome that would not be affirmed by any of the voters. Moreover, a judge's inconsistency will necessarily be between some belief about evidence and some belief about the consequences of the evidence — since the evidence claims are stipulated to be independent — but *premise-majority* does not query their beliefs about consequences at all. So it will generate a consistent set of judgements even if many or all judges are inconsistent. As such, *premise-majority* will generate determinate results even when the community is confused or fractured into competing camps. But, in considering scientific opinion, we certainly only want to say that there is something 'scientists know' when there is a coherent scientific community.

Second, applying the rule requires a division between the judgements that are evidence and the ones that are conclusions. As Fabrizio Cariana notes, *premise-majority* "requires us to isolate, for each issue, a distinguished set

[1]The strategy of adding complications only as necessary can be applied generally to decision problems. For example, intransitive preferences wreck dominance reasoning. Yet one might presumptively employ dominance reasoning until one actually faces a case where there are intransitive preferences.

[2]Since Hartmann et al. are thinking about the general problem of judgement aggregation, rather than the problem of expert elicitation, these are objections to the application of the rule rather than to the rule *as such*.

of logically independent premises" [4, p. 28]. He constructs a case involving three separate, contentious claims and an agreed upon constraint, such that any two of the three claims logically determines the third. It would be arbitrary to treat two of the claims as evidence (and so suitable for polling) and the third as a consequence (and so fixed by inference). The *premise-majority* rule simply is not applicable in cases where the line between premises and conclusions is so fluid. This difficulty leads Cariani to conclude only that *premise-majority* will sometimes be inapplicable; so he suggests, "Different specific aggregation problems may call for different aggregation rules" [4, p. 29]. Yet the problem is especially acute for scientific judgement, because inference can be parsed at different levels. Individual measurements like '35° at 1:07 AM' are not the sort of thing that would appear in a scientific publication; individual data points are unrepeatable and not something about which you would query the whole community. Yet they do, of course, play a rôle in inference. At the same time, scientists may take things like the constancy of the speed of light to be evidence for a theory; the evidence here is itself an inference from experiments and observations. There are different labels for these different levels. Trevor Pinch [12] calls them observations of differing *externality*. James Bogen and James Woodward [2] distinguish *data* from *phenomena*. Since we might treat the same claims as premises or conclusions, in different contexts, it is unclear what we would poll scientists about if we applied *premise-majority*.

Third, *premise-majority* is constructed for cases where the conclusion is a deductive consequence of the premises. In science, this is almost never the case.[3] Scientific inference is ampliative, and there is uncertainty not only about which evidence statements to accept but also about which inferences ought to be made on their basis. One might avoid this difficulty by including inferential relations among the evidential judgements. To take a schematic case, judges could be asked about R and $(R \to S)$; if the majority affirms both, then *premise-majority* yields an affirmative judgement for S.

One might worry that this suggestion treats ampliative, scientific inference too much like deductive consequence. The worry is that actual scientists might accept a premise of the form 'If R, then typically S' but nothing so strong as $R \to S$. It is possible for inferences based on weaker conditionals

[3] I say 'almost' because sufficiently strong background commitments can transform an ampliative inference into a deduction from phenomena. Of course, we accept equivalent inductive risk when we adopt the background commitments; cf. [9].

about what is merely typical to lead from consistent premises to inconsistent conclusions. To answer the worry, one might appeal to what John Norton [11] calls a *material theory of induction*. The central idea is that most of inductive risk in ampliative inferences is shouldered by conditional premises; Norton calls the premises *material postulates*. So — in answer to the worry — one might think that asking about material postulates would allow us to use the *premise-majority* rule to aggregate scientific judgements about many even though not absolutely all matters.

A deeper problem with the suggestion is that it presumes that scientists can say, independently of everything else, whether the inference from R to S is appropriate. That is, it assumes that material postulates can be evaluated on a ballot separately from everything else. In the remainder of the paper, I argue that this idealizes science too much. Whether a scientific inference is appropriate must be informed by *more* than just the particular evidence — the appropriate scientific conclusion depends (at least in some cases) on the risks and values involved.

In the next section, I spell out more clearly the way in which inference can be entangled with values and risk. In the subsequent section, I return to it as a problem for *premise-majority*. As we'll see, it becomes a problem for more than just Hartmann et al.'s specific proposal. It is a problem for any formal judgement aggregation rule whatsoever.

2 The James-Rudner-Douglas thesis

Here is a quick argument for the entanglement of judgement and values: There is a tension between different epistemic duties. The appropriate balance between these duties is a matter of value commitments rather than a matter of transcendent rationality. So making a judgement of fact necessarily depends on value commitments.

The argument goes back at least to William James, who puts the point this way: "*We must know the truth;* and *we must avoid error* — these are our first and great commandments as would-be knowers; but they are not two ways of stating an identical commandment, they are two separable laws" [8, p. 99]. Although James has in mind personal matters of conscience (such as religious belief), Richard Rudner makes a similar argument for scientific judgement. Rudner argues that

> the scientist must make the decision that the evidence is *suf-*

> *ficiently* strong... to warrant the acceptance of the hypothesis. Obviously our decision regarding the evidence and respecting how strong is "strong enough", is going to be a function of the *importance*, in the typically ethical sense, of making a mistake in accepting or rejecting the hypothesis. [13, p. 2]

There is not only a tension between finding truth and avoiding error, but also between risking one kind of error and risking another. Any particular test involves a trade-off between making the standards too permissive (and so mistakenly giving a positive answer) or making them too strict (and so mistakenly giving a negative answer). The former mistake is a *false positive* or *type I* error; the latter a *false negative* or *type II* error. There is an inevitable tradeoff between the risk of each mistake, and so there is a point at which the only way to reduce the risk of *both* is to collect more evidence and perform more tests. Yet the decision to do so is itself a practical as well as an epistemic decision. In any case, it leaves the realm of judgement aggregation — having more evidence would mean having different science, rather than discerning the best answer our present science has to a question. As such, values come into play. Heather Douglas puts the point this way, "Within the parameters of available resources and methods, some choices must be made, and that choice should weigh the costs of false positives versus false negatives. Weighing these costs legitimately involves social, ethical, and cognitive values" [5, p. 104].

Plotting a curve through these 19th, 20th, and 21st-century formulations, call this the *James-Rudner-Douglas* or *JRD thesis*: Anytime a scientist announces a judgement of fact, they are making a tradeoff between the risk of different kinds of error. This balancing act depends on the costs of each kind of error, so scientific judgement involves assessments of the value of different outcomes.

The standard objection to the thesis is that responsible scientists should not be making categorical judgements. They should never simply announce 'P' (the objection says) but instead should say things like 'The available evidence justifies $x\%$ confidence in P.' This response fails to undercut the thesis, because procedures for assigning confidence levels also involve a balance between different kinds of risk. This is clearest if the confidence is given as an interval, like $x \pm e\%$. Error can be avoided, at the cost of precision, by making e very large. Yet a tremendous interval, although safe, is tantamount to no answer at all.

Eric Winsberg and Justin Biddle [1] give a substantially more subtle reply to the standard objection. Regarding the specific case of climate modeling, Winsberg and Biddle show that scientists' estimates both of particular quantities and of confidence intervals depend on the histories of their models. For example, the results are different if scientists model ocean dynamics and then add a module for ice formation rather than vice-versa. The history of a model reflects decisions about what was considered to be important enough to model first, and so it depends on prior value judgements.

But why should the JRD thesis have consequences for expert elicitation? After all, James does not apply it to empirical scientific matters. He is concerned with religious and personal matters, and he concludes merely that we should "respect one another's mental freedom" [8, p. 109]. He does not apply it at all scientific matters where there is a community of legitimate experts.

Rudner, who does apply the thesis to empirical judgements, nevertheless hopes that the requisite values might themselves be objective. What we need, he concludes, is "a science of ethics" [13, p. 6]. Rudner calls this a "task of stupendous magnitude" [13, p. 6], but he is too optimistic. Searching for an objective ethics in order to resolve the weight of values and risks is a fool's errand. A regress would ensue: The judgements of ethical science would need to be informed by the ethically correct values so as to properly balance inductive risks, but assurance that we have the correct values would only be available as the product of ethical science. One might invoke pragmatism and reflective equilibrium, but such invocations would not give Rudner final or utterly objective values. If responsible judgement aggregation were to wait on an utterly objective, scientific ethics, then it would wait forever.

Douglas accepts that the thesis matters for expert elicitation. So she considers the concrete question of how to determine the importance of the relevant dangers. She argues for an *analytic-deliberative process* which would include both scientists and stakeholders [5, ch. 8]. Such a process is required when the scientific question has a bearing on public policy, and there are further conditions which must obtain in order for such processes to be successful. For one, "policymakers [must be] fully committed to taking seriously the public input and advice they receive and to be guided by the results of such deliberation" [5, p. 166]. For another, the public must be "engaged and manageable in size, so that stakeholders can be identified and involved" [5, p. 166]. Where there are too many stakeholders and scientists for direct interaction, there can still be vigorous public examination of the values

involved. Rather than pretending that there is any all-purpose procedure, Douglas calls for "experiment with social mechanisms to achieve a robust dialog and potential consensus about values" [5, p. 169]. Where consensus is impossible, we can still try to elucidate and narrow the range of options. Douglas' approach is both a matter of policy (trying to increase trust in science, rather than alienating policymakers and stakeholders) and a matter of normative politics (claiming that stakeholders' values are ones that scientists should take into consideration). In cases where these concerns are salient, saying *what scientists know* will depend on more than just the prior isolated judgements of scientists — but moreover on facts about the actual communities of scientists, policymakers, and stakeholders.

Arguably, Douglas' concerns will not be salient in all cases. Some science is far removed from questions of policy. So the significance of the JRD thesis may depend on the question being asked.

3 Our fallible selves

I argued above that the *premise-majority* rule was inapplicable in many scientific contexts because it only worked for cases of deductive consequence. Formally, this worry could be resolved by asking scientists about which inferences would be justified; we poll them about claims like $(E \to H)$ at the same time as we poll them about E. The JRD thesis undercuts this formal trick. Where the judgement has consequences, the inference itself is an action under uncertainty. So the appropriate inference depends on the values at stake. Schematically, whether one should assent to $(E \to H)$ depends on the risks involved in inferring H from E. Concretely, questions of science that matter for policy are not entirely separable from questions of the policy implications.

If we merely poll scientists, then we will be accepting whatever judgements accord with their unstated values. We instead want the procedure to reflect the *right* values, which in a democratic society means including communities effected by the science. Importantly, this does not mean that stakeholders get to decide matters of fact themselves; they merely help determine how the risks involved in reaching a judgement should be weighed. Nor does it mean that politicized scientific questions should be answered by political means; climate scientists can confidently identify general trends and connections, even allowing for disagreement about the values involved. What

it does mean is that scientists cannot provide an account that is value-neutral in all its precise details.[4]

This is fatal to *premise-majority* as a method of determining what scientists know collectively. Moreover, it is fatal to any judgement aggregation rule that treats judges merely as separate inputs to an algorithm. The problem extends to practical policies of expert elicitation, insofar as they are procedures for enacting judgement aggregation rules. Where there are important values at stake that scientists are not taking into account or where the value commitments of scientists are different than those of stakeholders, the present judgements of individual scientists can not just be taken as givens.

An analytic-deliberative process is required, but the appropriate mechanisms are not ones which we can derive *a priori*. As Douglas argues, we need to experiment with different possibilities [5, p. 169, cited above]. There is not likely to be one universally applicable process. It will depend on facts about the communities involved. Moreover, the inference from social experiments in deliberation will itself be an inductive inference about a question that effects policy. So the inference depends importantly on value judgements about the inductive risks involved, and that means an analytic-deliberative process will be required. It would be a mistake to hope, in parallel with Rudner's appeal to a science of ethics, for an objective set of procedural norms. How best to resolve meta-level judgement about experiments in social arrangements is as much a contingent matter as how to socially arrange object-level expert consultation. We start with the best processes we can muster up now, and we try to improve them going forward. Minimally, however, we can say that future improvements should not elide the rôle of values, as formal judgment aggregation functions do, but explicitly accommodate it.

References

[1] Justin Biddle and Eric Winsberg. Value judgements and the estimation of uncertainty in climate modelling. In P.D. Magnus and Jacob Busch, editors, *New Waves in Philosophy of Science*, pages 172–197. Palgrave MacMillan, Basingstoke, Hampshire, 2010.

[4]Douglas [5, esp. ch. 6] provides an excellent discussion of how (what I have called) the JRD thesis is compatible with objectivity.

[2] James Bogen and James Woodward. Saving the phenomena. *Philosophy of Science*, 97(3):303–352, July 1988.

[3] Steven J. Brams, D. Marc Kilgour, and William S. Zwicker. The paradox of multiple elections. *Social Choice and Welfare*, 15(2):211–236, 1998.

[4] Fabrizio Cariani. Judgment aggregation. *Philosophy Compass*, 6(1):22–32, 2011. doi:10.1111/j.1747-9991.2010.00366.x.

[5] Heather E. Douglas. *Science, Policy, and the Value-free Ideal*. University of Pittsburgh Press, 2009.

[6] Stephan Hartmann, Gabriella Pigozzi, and Jan Sprenger. Reliable methods of judgment aggregation. *Journal for Logic and Computation*, 20:603–617, 2010.

[7] Stephan Hartmann and Jan Sprenger. Judgment aggregation and the problem of tracking the truth. *Synthese*, forthcoming.

[8] William James. The will to believe. In Alburey Castell, editor, *Essays in Pragmatism*, pages 88–109. Hafner Publishing Co., New York, 1948. Originally published June, 1896.

[9] P.D. Magnus. Demonstrative induction and the skeleton of inference. *International Studies in the Philosophy of Science*, 22(3):303–315, October 2008. doi:10.1080/02698590802567373.

[10] Lynn Hankinson Nelson. *Who Knows: From Quine to a Feminist Empiricism*. Temple University Press, Philadelphia, 1990.

[11] John D. Norton. A material theory of induction. *Philosophy of Science*, 70(4):647–670, October 2003.

[12] Trevor Pinch. Towards an analysis of scientific observation: The externality and evidential significance of observational reports in physics. *Social Studies of Science*, 15:3–36, 1985.

[13] Richard Rudner. The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1):1–6, January 1953.

Is race a cause?

Alexandre Marcellesi

Draft of 11/03/2012
To be published in revised form in *Philosophy of Science*
Comments welcome (amarcellesi@ucsd.edu)

Abstract

Advocates of the counterfactual approach to causal inference argue that race isn't a cause. I object that their argument is invalid and that its key premise is unwarranted. I also criticize the criterion, which I call 'Holland's rule', the counterfactual approach relies on to distinguish causes from non-causes. Finally, I argue that racial discrimination cannot be causally explained unless one assumes race to be a cause. I conclude that the view that race is not a cause lacks support and that there are good reasons to adopt the opposite view that race is a cause.

1 Introduction

Scientists in many disciplines (economics, epidemiology, etc.) routinely treat race as a cause. Economists who study labor market discrimination, for instance, commonly build models involving race as an independent variable and give a causal interpretation of the coefficient attached to it.[1]

Are scientists who treat race as a cause fundamentally confused? Do policies based on their conclusions rest on shoddy evidence? This is what leading advocates of the counterfactual approach to causal inference (henceforth 'CFA') claim, arguing that since race is an "immutable characteristic" of individuals, one cannot coherently treat it as a cause.

After a brief introduction to the CFA (§2), I present the argument against race being a cause (§3). I then raise two objections to it (§4) and proceed to sketch a positive argument for race being a cause (§5). I conclude that advocates of the CFA lack justification for denying race the status of cause, and that there are good reasons to adopt the opposite view that race is a cause (§6).

2 The counterfactual approach

The CFA, first introduced by Rubin (1974), is the dominant approach to causal inference in statistics and in many social sciences. It has roots in the work of Fisher and Neyman on agricultural experiments.

[1] See e.g. (Kahn and Sherer, 1988) for a classic example that is representative of many studies in labor economics.

When only one cause is considered, counterfactual causal models essentially have the following components:[2]

- A population of units $i \in U$
- A binary causal exposure variable D taking value $d_i = 1$ when i is exposed to the cause (is in the 'treatment' state) and $d_i = 0$ when i is not (is in the 'control' state).
- Two potential outcome variables Y^1 and Y^0, where y_i^1 represents the value of the effect for i when i is exposed to the cause and y_i^0, the value of the effect for i when i is not exposed to the cause.

The individual-level causal effect (ICE) of D for i is typically defined as follows:

$$\delta_i = y_i^1 - y_i^0$$

This causal effect is equal to the difference between the value of the effect when i is exposed to the cause and the value of the effect when i is not. Since a given unit cannot be both exposed to the cause and not exposed to it at once, only one of y_i^1 and y_i^0 can be observed for any unit. If i is exposed to the cause, the value of y_i^1 is observable while the value of y_i^0 is counterfactual: It is the value the effect *would* have taken had i not been exposed to the cause; hence the name of the approach. Because only one of y_i^1 and y_i^0 can be observed, δ_i cannot be observed either. Holland dubs this the "fundamental problem of causal inference" (1986, 947).

There are various solutions to this problem, both in experimental and in observational contexts. These solutions provide techniques for estimating the ICE and other causal effects, or parameters, built upon it. My concern here is not with the problems that race might raise for the application of these estimation techniques.[3] It is, rather, with the problems that race allegedly raises for the very definition of causal effects, and of the ICE in particular.

3 The argument against race being a cause

The argument developed by leading advocates of the CFA against race being a cause can be reconstructed as follows:

1. Race is a necessary property of units
2. If a unit is of race r, then it is impossible for it to have been of another race r' (from 1)
3. Counterfactuals of the form 'Had i been of race r' instead of r, then…' cannot be (non-vacuously) true (from 2).

[2] I adopt the terminology and notation from (Morgan and Winship, 2007).

[3] Rubin (1986; 2011) argues that estimating the causal effects of race is difficult enough to warrant its dismissal as a cause. I agree with Heckman (2005) that arguments of this kind conflate definition and estimation: That it is difficult to estimate the causal effect of race does not warrant the conclusion that it is not a cause.

4. The ICE of race is undefined (from 3 and the definition of ICE).

$\therefore$ Race is not a cause (from 4).

Let me illustrate this argument. Assume that there are only two races, that D represents race, and that $d_i = 1$ when i is White and $d_i = 0$ when i is Black. To say that race is a necessary property, "immutable characteristic" (Greiner and Rubin, 2011), or "attribute" (Holland, 1986, 955) of units is to say that if $d_i = 1$ (resp. 0), then it could not have been the case that $d_i = 0$ (resp. 1). Because this is so, counterfactuals of the form 'Had it been the case that $d_i = 0$ instead of $d_i = 1$, then the value of Y^0 for i would have been y_i^0' cannot be non-vacuously true when $d_i = 1$. Because no such counterfactual can be non-vacuously true, the causal effect of race is undefined, and this regardless of what effect the variables Y^1 and Y^0 represent (wages, education, etc.).[4]

In Holland's words, "attributes of units [like race] are not the types of variables that lend themselves to *plausible states* of counterfactuality." (2003, 14, emphasis original) He adds: "Because I am a White person, it would be close to ridiculous to ask what would have happened to me had I been Black." (ibid.) And because the causal effect of race cannot be defined unless there is a non-vacuously true answer to such a counterfactual query, Holland concludes that race is not a cause.

The consequences of this view are important. If race is not a cause, then as Greiner and Rubin point out, "attempts to infer the causal effects of such traits [as race] are incoherent." (2011, 775) Holland goes further by claiming that, "Attributing cause to RACE is merely confusing and unhelpful in an area where scientific study is already difficult" and that, "Obscuring [the topics of discrimination and bias] with simplistic calculations that do not attend to the proper role of RACE in a causal study helps no one." (2003, 24)

So, do the many scientists who treat race as a cause waste time and resources on incoherent studies that only obscure important topics like racial discrimination? I do not believe so and develop two objections to the argument against race being a cause.

4 Against the argument against race being a cause

4.1 The argument is invalid

The most straightforward objection to the argument presented in §2 is that its conclusion does not follow from its premises. The fact that the ICE of race is undefined only entails that race is not a cause if the following premise is added to the argument:

4′. For all x, if x is a cause, then its ICE is defined.

[4]The same point applies mutatis mutandis to other causal effects defined in the CFA, e.g. the average causal effect defined over U as $E[Y^1] - E[Y^0]$. Because ICE is the fundamental causal effect for the CFA, however, I focus on it in the present paper.

If one adds this premise, then the argument is valid. There are good reasons, however, to believe that this premise is false, i.e. there are good reasons to think that some genuine causes cannot be handled by the CFA.

Holland himself claims, for instance, that scholastic achievement in primary school cannot be treated as a cause of the choice of secondary school by the CFA (1986, 955). Setting aside the question of the justification for this claim, the right conclusion to draw here is not that scholastic achievement is not a cause of school choice: There are good reasons to think that how well a student does in primary school has an effect on what secondary school she chooses to attend (e.g. by determining what schools she's admitted to). Rather, the conclusion to draw is that some genuine causes cannot be handled by the CFA, and so that premise $4'$ is false.

This conclusion is bolstered by the existence of frameworks for causal inference, e.g. Ragin's qualitative comparative analysis framework (1987), that do not rely on counterfactuals to define causal effects and which can thus treat variables whose ICE is undefined as causes.

4.2 Why believe premise 1?

How do advocates of the CFA justify the claim that race is an attribute, i.e. a necessary property, of units?[5] Their justification for this claim derives entirely from an application of what I will call 'Holland's rule' (or 'HR'). According to HR,

> If the variable *could be* a treatment in an experiment (even one that might be impossible to actually pull off due to ethical or practical issues), then the variable is [...] correctly called a *causal variable*. (2003, 9, emphasis original)

It is important to note that, for Holland, attributes and causal variables form a partition of the set of properties of a unit: If a property is not a causal variable, then it is an attribute. Holland claims that race could not be a treatment in an experiment and, applying HR, he thus concludes that it is not a causal variable but, rather, an attribute (ibid.).[6] Greiner and Rubin agree and invoke "the impossibility of manipulating such traits [as race] in a way analogous to administering a treatment in a randomized experiment" (2011, 775) as one of the sources of the incoherence of studies purporting to estimate the effect of race.

There are two important problems with HR. First, it is the wrong rule for advocates of the CFA to follow. According to the CFA, for the ICE of D on i to be defined, there must be some counterfactual state in which i is not exposed to D, assuming that i actually was exposed to D.

[5] Glymour has objected to Holland that, "If counterparts [in the sense of (Lewis, 1968)] are conceivable – and why not? – then counterfactuals that violate identity conditions are intelligible, and if counterfactuals are intelligible, then causal relations are as well." (Glymour, 1986) Holland, however, can answer this objection by saying that the problem with attributes is not that they engender counterfactuals which violate identity conditions, but that they engender counterfactuals with impossible antecedents. In other words, Holland could answer that though counterparts are conceivable, no counterpart of a White unit can be, e.g., Black. Because race is a necessary property, all counterparts of a White unit also are White.

[6] Note that Holland's argument is fallacious given the way HR is formulated: It denies the antecedent of HR and infers the negation of its consequent. I'm here adopting a charitable reading according to which it is *necessary* for a property of units to be a causal variable that it be a treatment in some possible experiment.

In other words, it must be possible for i not to have been exposed to D. But why think that the possibility of such a state requires the possibility of an experiment resulting in it being the case that i is not exposed to D? To hold this view is to hold the implausible view that it is possible that p only if it is possible for there to be an experiment of the right kind resulting in it being the case that p. The right slogan for the CFA thus isn't "No causation without [some hypothetical experimental] manipulation" (Holland, 1986, 959) but, rather, 'No causation without counterfactual states'. This slogan is less catchy but more faithful to the way the CFA defines causal effects (e.g. the ICE).

One might object that HR was intended by Holland not as a strict rule but as a heuristic. It is true that he prefaces his presentation of HR by saying that, "There is no cut-and-dried rule for deciding which variables in a study are causal and which are not." (2003, 9) It should be noted, however, that despite this caveat, Holland *does* apply HR as a "cut-and-dried" rule, since he takes the supposed violation of HR by race to be sufficient to establish the conclusion that race is an attribute and so cannot be a cause (op. cit., 10). It should also be noted that HR fares no better as a heuristic rule than it does as a strict rule. I have claimed above that the possibility of an experiment resulting in i not being exposed to D is not necessary for it to be possible that i is not exposed to D. If so, however, then there is no reason to take the inconceivability of such an experiment to be a reliable guide to the impossibility of a state in which i is not exposed to D.

The second issue is that HR is vague – What kinds of experiments are admissible? What does it mean to say that a variable *could be* a treatment in an experiment? – and that, as a result, it is unclear that it is genuinely impossible for there to be an experiment in which race is the treatment. Indeed, let me argue against this impossibility claim by describing a hypothetical randomized experiment in which race is the treatment:[7] Assume that the race r_i of i is a function $r_i = f(b_i, e_i)$ of biological (b_i) and environmental (including social and cultural) factors (e_i).[8] Imagine that values of b_i and e_i, and thus also of r_i, are randomly assigned to embryos 30 days after conception. The biological factors are assigned via genetic engineering and the environmental factors are assigned by swapping embryos between mothers.[9]

This experiment has not been carried out, is morally objectionable, and *might* be practically impossible given present science and technology. But this does not mean that this experiment is impossible. Indeed, the experiment described seems to be nomologically possible, i.e. carrying it out would not seem to require the violation of any laws of nature. This experiment also clarifies what the antecedents of counterfactuals of the form 'Had i been of race r' instead of r, then...' claim. The race of i would have been different just in case i had been randomly assigned a combination of values of b_i and e_i giving rise to a value r' of r_i that differs from its actual value r.

[7]Note that HR does not require the relevant hypothetical experiments to be randomized. I am offering more than is required here.

[8]What the relative weights of b_i and e_i are is no concern of mine. If you think that race is entirely determined by biological factors, then give zero weight to e_i; and if you think that race is entirely determined by environmental (including social) factors, then give zero weight to b_i.

[9]Note that this experiment will not work if, among the biological factors represented by b_i, are 'genealogical' properties of i (e.g. who i's parents are). Thus, if you think that races are biological groups unified by genealogical relations (see e.g. Hardimon 2012), then you should think that the experiment described above does not randomly assign race.

It thus seems that, despite what Holland and Greiner and Rubin assume, it is possible for race to be a treatment in an experiment, even a randomized experiment, and so it is not the case that race violates HR. Even if HR was the right rule for advocates of the CFA to follow (a view I have argued against), then, its application to the case of race would not support the claim that race is an attribute of units rather than a causal variable, i.e. it would not support premise 1. So, not only is the argument against race being a cause invalid, as I have argued in §4.1, but its key premise also lacks proper support.

5 A positive argument for race being a cause: Explaining racial discrimination

Consider an imaginary society in which there are two exclusive and exhaustive racial groups, A and B. Assume that in this society there is a wage gap between As and Bs: As receive wages that are uniformly 30% lower than the wages received by Bs occupying equivalent jobs. Assume, further, that all the units in the population, be they A or B, are perfectly homogeneous regarding the causes of wages (other than, possibly, race), e.g. they received the same degree from the same school, they have the same number of years of experience, they have the same IT skills, they have the same interpersonal skills, they work equally hard, they have the same preferences regarding wages, etc. Assume, finally, that there is only one employer in this society, and that this employer fixes the wages of the workers hired.

What is the mechanism generating this wage gap, i.e. what causally explains the fact that As receive wages that are 30% lower than those of Bs? The seemingly obvious answer is that As receive lower wages precisely because they are As and because the employer believes that the work of As is worth 30% less than that of Bs. In other words, what causes As to receive lower wages is the fact that they are As combined with the fact that the employer believes the work of As to be worth 30% less than that of Bs.

This commonsensical causal explanation is unavailable to somebody who claims that race is not a cause. If being an A is not a cause, then being an A cannot, in combination with the employer's belief about the worth of the work of As, cause one to receive lower wages. But, then, what causally explains the wage gap between As and Bs? Let me consider two alternative ways one might answer this question below.[10]

The first alternative answer, defended by Holland (2003), consists in claiming that what causally explains the wage gap is not the racial difference between As and Bs but, rather, the (racially) discriminatory nature of the society I described. This answer, however, faces an immediate difficulty. For advocates of the CFA, 'being discriminatory' must satisfy HR in order for the discriminatory nature of society to be a cause of the wage gap. In other words, it must be possible for 'being discriminatory' to be assigned as a treatment to societies in some experiment. Is such an experiment

[10] I leave aside two implausible solutions: First, the solution which consists in claiming that the wage gap is a brute fact, i.e. has no causal explanation. Second, the solution which consists in claiming that a society of the kind I've described is impossible, and that As and Bs must differ in some respect other than their race in order for the wage gap to arise.

possible?

Holland attempts to justify his claim that it is by describing "a *parallel world* [...] in which things are so different that what we recognize in our own world as racial discrimination does not exist in this other world." (2003, 16, emphasis original) Though Holland attempts to further flesh out this "little fantasy" (ibid.), his description falls far short of a precise description of a hypothetical experiment. He does not specify, for instance, the hypothetical experimental manipulations involved in making a society discriminatory.[11]

The claim that 'being discriminatory' satisfies HR, and so might be a cause of the wage gap between As and Bs, thus lacks proper justification while, as I argued in §4.2, there are good reasons to think that race does satisfy HR. If HR is the right rule for advocates of the CFA to follow, then, there does not seem to be any good reason to favor Holland's alternative explanation of the wage gap over the commonsensical explanation I presented above. And if, as I argued in §4.1, HR is not the right rule for advocates of the CFA to follow, then one can simply object to Holland that, absent an account of what it means exactly for a society to be discriminatory, his proposed explanation is little more than a vague suggestion while the commonsensical explanation given above clearly identifies a mechanism that is sufficient to generate the wage gap between As and Bs. Whether HR is the right rule for advocates of the CFA to follow, then, there are good reasons to favor the commonsensical explanation over Holland's alternative.

The second alternative answer, defended by Greiner and Rubin (2011), among others, claims that what causes As to receive lower wages is not their race in combination with the employer's belief regarding the worth of their work, but the perception of their race by the employer in combination with this same belief. There are several problems with this answer. I here examine three.

First, in the imaginary case at hand, it is simple enough to pin down who's perception it is that's relevant to explaining the wage gap, since there is only one employer. But what if there were many employers, and what if the wages of As were on average, rather than uniformly, 30% lower than those of Bs? Who's perception would then be relevant? The collective perception of all the employers? Or the collective perception of only those employers who falsely believe the work of As to be worth less than that of Bs? If one is to appeal to perceptions of race to explain any real wage gap between racial groups, then one needs answers to these questions. Greiner and Rubin themselves point out the difficulty of answering these questions as one limitation of this approach (ibid., 783-784). And the problem is more severe even when one considers studies of the effect of race on education or access to health care: What is the proper interpretation in terms of perceptions of race of the causal effects estimated by these studies?

Second, if the move to perceptions is legitimate in the case of race, then why not adopt it for other properties of units? Why not think that *perceptions* of education or work experience, rather than education or work experience, are what's causally relevant to an individual's wages? The move from race to perceptions of race seems ad hoc and motivated entirely by the assumption, which I

[11] It seems that, in order to describe such hypothetical experimental manipulations, one would first have to pin down what it means for a society to be (racially) discriminatory, something Holland does not do.

have argued to be mistaken in §4, that race cannot be a cause according to the CFA.

Third, and this is the most pressing problem, what causes the employer in the imaginary society I have described to perceive A workers to be As? If race is not a cause, then one cannot claim that the cause at work is the fact that As are of race A. Leaving aside the implausible claim that perceptions of race are uncaused, the most plausible solution seems to be to claim that what causes the employer to perceive A workers to be As is the perception of some set of features F the presence of which is strongly correlated with being a A. This solution faces a dilemma. Either the features in set F constitute what it is to be an A, in which case being an A is, after all, a cause of the employer's perception of As as As. Or the features in F do not constitute what it means to be a A.

In this latter case, the belief the employer must have in order for the wage gap to appear is not the belief that the work of As is worth less than that of Bs, but the belief that the work of units exemplifying features F is worth less than the work of units which do not exemplify these features. If this is so, then describing the discrimination against As as *racial* discrimination is inappropriate: As are discriminated against not on the basis of their race but on the basis of features that happen to be strongly correlated with being an A. More generally, this solution amounts to denying that there can be genuinely *racial* (direct) discrimination, i.e. "*differential treatment on the basis of race* that disadvantages a racial group", as a panel of the US National Research Council defines it (Blank et al., 2004, 39, emphasis original).

Neither the alternative explanation defended by Holland nor that defended by Greiner and Rubin thus seem as satisfactory as the commonsensical explanation presented at the beginning of this section, and which assumes race to be among the causes of the wage gap between As and Bs. This provides some support for the claim that one needs to assume race to be a cause in order to explain racial discrimination. Of course, the explanations offered by Holland and Greiner and Rubin, though they are the most prominent alternatives in the literature (especially the latter), do not exhaust the space of possible alternatives to the commonsensical explanation of the wage gap. This is why, in the introduction to this paper, I described the discussion in the present section as *sketching* an argument for race being a cause.

6 Conclusion

Are the attempts of labor economists to infer the causal effect of race on, e.g., wages "incoherent", as Greiner and Rubin (2011, 775) claim? Is it the case that "Attributing cause to RACE is merely confusing and unhelpful", as Holland (2003, 24) maintains? I have here argued that there is no reason to think these claims to be true.

First, the argument advanced by advocates of the CFA against race being a cause is invalid. Second, its key premise, that race is an attribute of units, is not justified by the application of Holland's rule, a rule that advocates of the CFA should reject anyway. Third, there are good reasons to think that explaining racial discrimination requires one to treat race as a cause.

I have said nothing up to now about debates in the philosophy of race. The view defended in this paper bears on these debates in the following way: Whatever concept of race one thinks is fit for use by labor economists studying racial discrimination, one's account of this concept should imply that races can be causes.

The debate over the causal status of race examined in this paper also gives a useful example of a case where philosophers of science can, and should, contribute to clarifying the debate and critically examine the assumption made by the scientists involved. This is what I have tried to do above.

Acknowledgments

I wish to thank Craig Callender, Nancy Cartwright, Michael Hardimon, Gil Hertsthen, and Chris Wüthrich for comments on earlier drafts. I also wish to thank members of the audience at the UCSD Graduate Philosophy Colloquium, in particular Amy Berg, Dan Burnston, Joyce Havstad, Nat Jacobs, Chris Pariso, and Adam Streed.

References

Blank, Rebecca, Dabady, Marilyn, and Citro, Constance (eds.). 2004. *Measuring Racial Discrimination*. Panel on Methods for Assessing Discrimination. Washington, D.C.: The National Academies Press.

Glymour, Clark. 1986. "Comment: Statistics and Metaphysics." *Journal of the American Statistical Association* 81:964–966.

Greiner, James and Rubin, Donald. 2011. "Causal effects of perceived immutable characteristics." *The Review of Economics and Statistics* 93:775–785.

Hardimon, Michael. 2012. "The Idea of Scientific Concept of Race." *Journal of Philosophical Research* 37:249–282.

Heckman, James. 2005. "Rejoinder: Response to Sobel." *Sociological Methodology* 35:135–150.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

—. 2003. "Causation and Race." Technical Report RR-03-03, Educational Testing Services.

Kahn, Lawrence and Sherer, Peter. 1988. "Racial Differences in Professional Basketball Players' Compensation." *Journal of Labor Economics* 6:40–61.

Lewis, David. 1968. "Counterpart Theory and Quantified Modal Logic." *Journal of Philosophy* 65:113–26.

Morgan, Stephen and Winship, Christopher. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

Ragin, Charles. 1987. *The Comparative Method*. University of California Press.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.

—. 1986. "Comment: Which ifs have causal answers?" *Journal of the American Statistical Association* 81:961–962.

Is the Contingentist/Inevitabilist Debate a Matter of Degrees?[*]

Joseph D. Martin[†]

† University of Minnesota, Program in the History of Science, Technology, and Medicine; Minnesota Center for Philosophy of Science; mart1901@umn.edu

Abstract: Debates between contingentists and inevitabilists contest whether the results of successful science are contingent or inevitable. This paper addresses lingering ambiguity in the way contingency is defined in these debates. I argue that contingency in science can be understood as a collection of distinct concepts, distinguished by how they hold science contingent, by what elements of science they hold contingent, and by what those elements are contingent upon. I present a preliminary taxonomy designed to characterize the full range positions available and illustrate that these constitute a diverse array, rather than a spectrum.

1. Introduction

Ian Hacking, in *The Social Construction of What?*, asks his readers to assign themselves a number from one to five to describe how central contingency is to their personal conceptions of science. If you rate yourself at one, then you are a strong inevitabilitst, whereas if you choose five, you are highly contingentist and probably have strong constructionist sympathies (Hacking 1999, 99). In response, Léna Soler questions whether this is the correct approach, and asks: "should we introduce degrees of contingentism depending on the kind of contingent factors that are supposed to play a role?" (Soler 2008a, 223).

Herein, I answer Soler's question in the emphatic affirmative, and therefore the question posed in the title with a resounding "no." Contingency in science can be understood as a collection of distinct concepts, distinguished by how they hold science contingent, by what elements of science they hold contingent, and by what those elements are contingent upon. What separates one contingentist from another is not that one tags herself a two and the other fancies himself a five according with how strongly each believes science might have developed

[*] This version accepted for presentation at PSA 2012, San Diego, CA. Final version forthcoming in *Philosophy of Science*.

differently. Their disagreement arises from the fact that they understand contingency-producing factors to act differently on different aspects of the scientific process. Contingency is a "what" question, not a "how much" question.

Before beginning this discussion I review the contingentist/inevitabilist (C/I) debate in Section 2 by reconstructing positions the debate's central figures stake out. Ian Hacking, who coined the terms "contingentism" and "inevitabilism," figures centrally. I also discuss several scholars who were retrospectively cast as interlocutors in the debate, such as Andrew Pickering, Sheldon Glashow, and James Cushing, and those who responded to Hacking directly, namely Léna Soler and Howard Sankey. After demonstrating how their conceptions of contingency have defined the debate, I argue that the conversation wants for a clear understanding of contingency and suggest how this ambiguity might be clarified by more rigorous classification of the concepts it groups together.

Section 3 presents a detailed discussion of the nature of contingency in science, in which I outline a fresh taxonomy of the concept. The taxonomy builds on John Beatty's distinction between unpredictability contingency and causal dependence contingency (Beatty 2006). This distinction clarifies the debate substantially, but I argue that a second step is required. Further decomposing unpredictability contingency and sub-classifying causal dependence contingency—based on the things within science considered to be contingent and the factors they are presumed to be contingent upon—allows more precise characterization of the views under discussion. A detailed picture of ways different authors use contingency serves as a basis from which to examine how a nuanced account of the concept can clarify some persistent ambiguities in the C/I debate.

2. Contingency and Inevitability

Ian Hacking coined "contingentism" and "inevitabilism" in the same book in which he hinted that contingency might be understood as a spectrum. Contingency appears as a feature of his effort to understand the philosophical stakes of social constructionism. Hacking casts contingency as a sticking points between constructionists and their opponents. He identifies the constructionist program as seeking to undermine claims about the inevitability of ideas. When generalized, according to Hacking, the constructionist argument takes the form "X need not have existed, or need not be at all as it is, is not determined by the nature of things; it is not inevitable." It often proceeds to two other more advanced stages, which contend a) that X is bad in its current form, and therefore b) should be eliminated or radically altered (Hacking 1999, 6). The constructionist program meets irreconcilable opposition from inevitabilists when it claims that the results of scientific investigation are contingent, and therefore unconstrained by the structure and properties of the natural world.

Andrew Pickering, author of 1995's *Constructing Quarks*, is Hacking's paradigm contingentist. Pickering advanced the view that high energy physics' Standard Model resulted from an exegesis of data, which could have produced any one of numerous, ontologically incompatible interpretations. He concludes that physics might have escaped the twentieth century quark free, and that if it had, it would not be any less successful (Pickering 1984). Hacking interprets this argument in light of later work, *The Mangle of Practice* (Pickering 1995), wherein Pickering argued that scientific consensus arises from negotiation between theory applied to the world, theory applied to instruments, and the construction of the instruments themselves to develop a robust fit with observed data. The results of science are contingent from this perspective because the negotiation could be carried out in any number of ways, each resulting in

the same degree of self-described success. Pickering's punch line is that twentieth-century physics could have been just as successful if, for example, cyclotrons had not supplanted traditional cloud-chamber technology and the resulting theory of the micro-world had not been dominated by quarks, which he contends are the peculiar progeny of the particle accelerator.

Hacking elaborates the inevitabilist stance in "How Inevitable Are the Results of Successful Science?," writing: "We ask: *If the results R of a scientific investigation are correct, would any investigation of roughly the same subject matter, if successful, at least implicitly contain or imply the same results?* If so, there is a significant sense in which the results are inevitable" (Hacking 2000, 61). Pickering would deny that equal success implies equivalence of any sort. By contrast, Hacking casts Sheldon Glashow as arch inevitabilist. Glashow holds that any investigation into the natural world starting from reasonable initial assumptions would produce not only the same answers, but also a similar set of questions to ask. Glashow imagines intelligent aliens as hypothetical scientists whose physical laws should be isomorphic with ours. In doing so, Hacking charges, Glashow tacitly makes crucial assumptions about the "reasonable" initial conditions necessary for alien science to produce the same results. How do we know, for example, that aliens would identify proton structure as an interesting question? Hacking segues from Glashow into the difficulties with strong inevitability claims: how stringently can you set the initial conditions before the argument dissolves into tautology? If the inevitabilist asserts that a successful alternate scientific enterprise will produce the same results by stipulating that success requires asking the same questions, using the same instruments to observe the same entities, and starting from the same assumptions, then we are left with the trivial observation that effectively identical scientific investigations produce effectively identical results (2000*,* 66).

Pickering and Glashow represent extremes; Hacking seeks a middle way. His compromise locates contingency at the level of the questions scientists ask. It is contingent, he argues, which questions are "live." Live questions are those that make sense within the contemporary theoretical framework. Once science satisfactorily answers a live question we can take that result to be inevitable in some meaningful sense, but we have no guarantee that it would have been asked in the first place.[1] Contingency, for Hacking, enters into science by allowing historical and socio-cultural factors to define what questions scientists find interesting and what questions they are allowed to ask. These questions are not necessarily answerable, and they might not make sense in any theory-independent sense, but once nature proves forthcoming with an answer, that answer has the tinge of inevitability. Science could have developed differently, but only because it could have addressed a different set of questions. Possible alternate results are never logically incompatible with current successful science (2000, 71).

When distinguishing contingency from inevitability, Hacking observes the debate's independence from the realism/anti-realism issue: "the contingency thesis itself is perfectly consistent with […] scientific realism, and indeed anti-realists […] might dislike the contingency thesis wholeheartedly," (Hacking 1999, 80). Howard Sankey (2008) maintains the same separation between the debates. He defends weak fallibilism, consistent with an inevitabilist viewpoint, holding that individual results of science are contingent—individual instances of scientific investigation are fallible—but we can be confident that statistically inevitabilist tendencies will wash out local contingencies.

Sankey defends his fallabilist stance's compatibility with a contingency thesis, which he says is an epistemic claim about scientific practice and the way investigators engage with the

[1] Hacking does not offer an account of just how scientists can determine when a live question has been adequately answered, an issue that is not unproblematic (see Galison 1987).

world: “Scientist might collect different evidence from the evidence they in fact do collect. They might have developed different instruments and techniques from the ones which have been developed and put to use” (Sankey 2008, 259). A geological example, the discovery of continental drift, illustrates his point: “The epistemic situation is […] dependent on contingent factors such as the availability of evidence and relevant knowledge, the development of instrumentation and the provision of research funding” (2008, 262). Sankey’s contingency differs from both Pickering’s and Hacking’s. Pickering would not contest that the factors Sankey identifies are contingent, but he would compile a list of additional contingencies much longer than Sankey would admit. Hacking argues for contingency of form rather than content of science: difference without incompatibility. Sankey points to the empirical content of science as contingent. These perspectives are not incompatible, but they have different emphases—Sankey focuses on evidence, Hacking on inquiry.

Sankey subtly contrasts James Cushing, who argues that contingency has an “ineliminable role in the construction and selection of a successful scientific theory from among its observationally equivalent and unrefuted competitors” (Cushing 1994, xi). Cushing uses “theory” equivocally, as his prime example is the choice between Bohr’s and Bohm’s interpretations of quantum mechanics, which can be construed as competing window dressings of the theory of quantum mechanics rather than as theories themselves. Quibbling aside, Cushing argues that choices between observationally equivalent theories are contingent. He does not claim that such choices are irrational, but that they are guided by philosophical and other external criteria. In the case of Bohm versus Bohr, the interpretive question hinges on whether one abandons strict determinacy or strict locality in the quantum realm. Evidence suggests that either particles in quantum states, obeying the probabilities assigned by their wave functions, assume

classically observable values for their key properties—charge, spin etc.—during an observation event, or some "hidden variables" determine these properties, but instantaneous signaling across finite distances is permitted. The first violates an ingrained philosophical preference for deterministic processes in physics, while the second flaunts a tradition of skepticism about instantaneous action at a distance. Cushing's view, exemplified by the claim that the Bohmian view's defeat at the hands of Bohr's Copenhagen interpretation was contingent, involves no change in the empirical content of the theories in question. Nor does Cushingtonian contingency act on the data collection process—the crux of Sankey's argument.

Most who deploy contingency do so in pursuit of goals other than defining it. Sankey wants to show the independence of the C/I debate from discussions of realism. Léna Soler identifies this argument as a premature, writing: "the 'contingentism versus inevitabilism' contrast does not exist as an autonomous, well identified issue of significance," (Soler 2008b, 232). On the basis of this ambiguity she sets out to clarify the issue, employing a thought experiment involving two, isolated communities of physicists, starting with the same initial conditions, asking their own questions, unguided by the work of the other scientists:

> Human beings might have succeeded in developing a physics as successful and progressive as ours, and yet asked completely different physical questions from the ones that have actually been asked, with the result that the accepted answers—in other words the content of the accepted physical theories and experimentally established physical facts—would be at the same time robust and different from ours. (2008b, 232)

Any non-trivial contingency, Soler contends, requires that two isolated scientific communities starting from the same point produce "irreducibly different" results, while still satisfying a reasonable set of criteria for success (2008b, 232).

Soler's contingency involves deep and irreconcilable oppositions between competing physical theories. Given the constancy of the initial conditions in Soler's thought experiment, it

tests only whether science is contingent irrespective of the initial conditions, and does not consider to what extent science might be contingent *upon* antecedent conditions.[2] Soler's thought experiment does not assess the relative contributions of contingency to the collection of internal and external factors that influence the trajectory of science.

Each scholar mentioned here questions how science might be contingent. In doing so, each employs a different understanding of what contingency means and at what point the claim becomes meaningful. They cast contingency in a qualitatively different ways rather than with differing intensities, representing diversity of kind, not of degree:

Hacking: *It is contingent what questions scientists decide are interesting.*

Pickering: *It is contingent what ontological entities scientists claim to find in the natural world.*

Glashow: *The theoretical structure of science is* ***not*** *contingent.*

Sankey: *It is contingent what instruments and techniques are available to scientists.*

Cushing: *It is contingent how scientists arbitrate between empirically equivalent theories.*

Soler: *Science is contingent only if it has available at least two equally successful, but irreducibly different paths from any given starting point.*

A smooth scale of contingentism cannot capture their differences, even superficially. The next section systematizes the diversity of views sheltered within the contingency concept.

3. Taxonomizing Contingency

3.1. A Preliminary Distinction

[2] Here I implicitly distinguish "contingent per se" from "contingent upon," borrowing from Beatty (2006). See Section 3 below for a more thoroughgoing discussion.

Contingency is a wildly diverse concept. How can we refine our understanding of contingency so it can be applied with less ambiguity? John Beatty offers a crucial distinction between "contingent per se" and "contingent upon" (Beatty 2006). "Per se" contingency describes stochasticity in the historical process; it implies that the process of history itself is *unpredictable*. "Upon" contingency requires no unpredictability, but rather describes a historical process that is far from robust with respect initial conditions, indicating that outcomes have a measure of *causal dependence* on the relevant antecedent factors. Any change in initial conditions could lead to a different outcome, even if the outcome of the process is, in principle, predictable from any given set of initial conditions.

In drawing this distinction, Beatty invokes Stephen J. Gould's thought experiment: restart the story of evolution from the Cambrian explosion, and ask if "replaying the tape" in this way directs the history of life down a different path (Gould 1989). Gould argues that evolution is highly contingent, and the rerun would differ dramatically from the initial broadcast. As Beatty observes, Gould alternates between the unpredictability and causal dependence senses of contingency. Beatty argues that these two conceptions are compatible, but have different consequences for our understanding of the historical process.

How should recognizing the distinction between these two varieties of contingency inform the C/I debate? Take Pickering: his 1984 claim that physics might have proceeded in a direction that did not include quarks is an unpredictability claim about scientific knowledge. He holds there that scientific knowledge is contingent per se. His view as reinterpreted by Hacking is an "upon" contingency claim. If the response to new data is a negotiation between existing theories, auxiliary theories about instruments, and the instruments themselves, then the consequent theory is contingent upon each of those three factors. In the second version of the

argument, Pickering's stance gets its bite from the factors it identifies as causally relevant rather than from the unpredictability of the scientific process.

Hacking, Soler, and Sankey, all observe that even the strongest inevitabilist admits that a benign form of historical contingency shapes the course of science. The Bragg family might have gone into sheep shearing rather than physics, and the resulting disturbance in the development of x-ray crystallography would likely have substantially altered the story of the discovery of DNA's structure. The Cold War might have dragged on a few years longer, the United States Congress might have been friendlier towards basic research expenditures, the Superconducting Super Collider might have been built, and high energy physicists might no longer be looking for the Higgs boson. In Beatty's language, inevitabilists are happy with the claim that scientific knowledge is contingent upon some historical factors, while denying the stronger claim that it is contingent per se.

Beatty's distinction substantially clarifies disagreements between inevitabilists and contingentists. They do not disagree about *the extent to which* scientific knowledge is contingent; they disagree about *what kind of* contingency influences the scientific process. Contingentists, as described by Hacking, admit both unpredictability and causal dependence contingency, while inevitabilists see no trouble from some types of causal dependence contingency, but draw the line at its more consequential sibling. This distinction does not exhaust the possible positions in the contingency debate. It demonstrates that Hacking's method of rating contingency on a spectrum inadequately describes the commitments involved, but it only begins to capture the full range contingency claims available. Those who allow causal dependence contingency might have reasonable disagreements about what aspects of science are subject to contingency claims and what science can be reasonably said to be contingent upon.

3.2. Towards a Taxonomy of Contingency

Each of Beatty's categories might be decomposed further. First, consider unpredictability contingency. Beatty defines it as the belief that "the occurrence of a particular prior state is *insufficient* to bring about a particular outcome," (Beatty 2006, 339). It appears that the unpredictability contingentist makes a strong metaphysical claim about the historical process: it is indeterministic. Indeed, Gould does appear to be making such indeterminacy claims. Should we replay the exact same tape of life from the exact same initial conditions and get a different result, then the process by which life develops exhibits intrinsic stochasticity.

Indeterminacy is not, however, the only way to understand per se contingency. Beatty observes that contingency is the lynchpin of Gould's argument that selection should not be the only causal agent evolutionary biologists invoke to explain the features and behaviors of present-day organisms (see Gould and Lewontin 1979). This suggests that unpredictability, as applied to contingency, can be understood as a methodological argument. This weaker understanding would suggest that outcomes are contingent (per se) *with respect to* some specified set of causal factors. It does not rule out the ability of other causal factors to provide an exhaustive, deterministic explanation. In fact, it often suggests such factors. Such is Gould's case against what he calls pan-selectionism—the assumption that selection can be invoked to explain any feature of an organism. The weaker version of unpredictability contingency he employs suggests that the features of organisms are contingent (unpredictable) *with respect to* selection effects. Such a view is consistent with deterministic evolution; it merely implies that factors other than selection are partly responsible.

The strong version of unpredictability contingency, which we might call indeterminist contingency, implies randomness in the historical process. The weaker version, incompleteness

contingency, claims that some set of causal factors is inadequate offer a complete explanation of the historical process, and that outcomes are unpredictable with respect to that set of factors. These two forms do different types of philosophical work. Indeterminist contingency says something about how the world is. Incompleteness contingency brands a set of explanatory tools inadequate, and so depends on the state of scientific practice and must refer to established explanatory orthodoxy.

Causal-dependence contingency is a more complicated case than unpredictability because the objects of "upon" might be expounded *ad nauseam*. The first step towards a classification requires identifying suitably distinct parts of science that might be held contingent. Science, like contingency, is heterogeneous and the claim that science is contingent can mean different things depending on what parts of science that claim specifies. Science makes ontological claims, formulates methodological procedures, develops models, adopts interpretations, and builds communities. Causal dependence contingency can be initially differentiated based on which of these many aspects of science are claimed contingent. I propose five categories:

(1) *Trivial contingency* – Science is part of a historical process, and so is contingent in the same way human history is contingent. This weak claim covers individual scientists and the details of their everyday existences.

All non-Laplacian parties are happy to admit this form of contingency. A claim that science is contingent in the trivial sense, however, offers the hard-boiled contingentist little succor. Trivial contingency is agnostic about the aspects of science that are typically of interest to philosophers, and so has little bearing on the debate. This type of contingency is frequently invoked to argue that contingency need not be repugnant to the sophisticated inevitabilist. Sankey, for instance, argues that continental drift did not gain traction within the geology

community until the 1950s and 1960s, when the U.S. Department of Naval Research began funding ocean floor research to bolster its submarine program (Sankey 2008, 262). Naturally, if the research had not been funded, and had not been conducted, the trajectory taken by the science would have been different, but this does not bear on the claim that successful science should pass through stages resembling ours. Trivial contingency alters the route science takes, but remains silent about its destination.

(2) *Sociocultural contingency* – The social structures that constitute scientific activity and science's interaction with culture are contingent.

At first glance this slightly stronger form of contingency might seem similarly innocuous. Like trivial contingency, it is agnostic about the content of science, acting instead on institutions, disciplines, communities, political relationships, and laboratory cultures. It is more complicated than trivial contingency, however, because it is the point where some strong contingentists dig in their heels. Forms of contingency that cut closer to the bone (see below) often rest on social determinism. A contingentist claiming that theoretical entities are contingent upon (causally determined by) social structures might want to deny that those social structures are themselves contingent. Similarly, inevitabilists might flinch when sociocultural contingency is used in conjunction with a stronger form, as in, for example, the controversial Forman thesis, which asserts that quantum indeterminacy was contingent upon the distinctive social conditions of the Weimar Republic (Forman 1971).

(3) *Methodological contingency* – The way in which we do science might have been different. This moderately weak variety holds experimental and theoretical techniques, laboratory practice, instruments, apparatus, and heuristic devices contingent.

Contingency claims frequently target the way science functions. Sankey approximates this version of contingency when he describes evidence collection and instrumentation as sources of contingency and claims that the development of plate tectonics could only come about when specific instrumentation came into common use (Sankey 2008). Many historical studies have examined how tool selection influences the way theories develop. The literature on model organisms is an obvious example. Robert Kohler's *Lords of the Fly* contends that the choice of *drosophila melanogaster* as the model organism for experimental genetics shaped the field's development (Kohler 1994). Experimental apparatus influences the collection, packaging, and inflection of data, while the available mathematics, heuristics, and analogies guide how that data is analyzed. This type of contingency is not trivial, but it does not directly imply incompatibilities between existing science and science that might have proceeded with different experimental or analytical tools. As with sociocultural contingency it can be combined with more potent forms.

(4) *Interpretive contingency* – The way in which we expound data in order to fill theoretical gaps is contingent.

Understanding theoretical implications requires interpreting data. Data, even if they motivate a particular theory, often do not compel one interpretation of that theory. Take Cushing's claim about the contingency of the Copenhagen interpretation: Quantum mechanics allows multiple logically consistent interpretations of what happens when quantum systems are observed. Building a satisfying ontological explanation requires physicists to interpret measurements that, by the very nature of the theory, do not provide the whole story. Given this necessary appeal to factors other than data, the interpretation we choose is contingent upon the

context in which the theory emerges, and an alternate interpretation might well have emerged given different conditions (Cushing 1994).

(5) *Theoretical contingency* – This is the strongest form of contingency. In the constructionist mold, it holds that scientific theories themselves and the claims they make about the world, are contingent.

This form postulates deep incompatibility between two possible scientific trajectories. While theoretical contingency can be parsed in "upon" syntax, it approximates a per se claim. The main difference between theoretical contingency and the in-principle unpredictability of scientific results is the frequent postulation by its advocates of a causal arrow from specific historical or cultural factors to theories. Forman's argument that cultural instability in the Weimar Republic compelled physicists to accept indeterminacy, for instance, makes quantum mechanics' ontological claims contingent upon the Weimar cultural environment (Forman 1971). This is not the same as describing science as unpredictable, but the factors on which it is contingent make the claim equivalent with the incompleteness contingency claim that science is unpredictable from internal factors alone. The per se claim and the theoretical contingency claim often go hand in hand, as the argument often holds that theoretical contingency works *because* theory is either almost infinitely malleable (indeterminist), and/or subject to pressures that are currently underappreciated (incompleteness).

It might appear that this constitutes a spectrum given a description beginning with "trivial" and graduating into increasingly more serious claims, but the relationships between the elements are not so straightforward. Trivial contingency does not require a commitment to any of the other four, and theoretical contingency often implies several of the others a fortiori, but middle-of-the-road contingency claims cannot be so easily ranked. It would be consistent to hold

an inevitabilist stance about methodology, arguing that mature science motivates an optimal form of investigation and modeling, while maintaining interpretive contingency. It would be equally consistent to be inevitabilist about interpretation while contingentist about methodology. These examples elucidate why contingency is a "what sort" question as opposed to a "how much" question. If I claim that one part of the scientific process is contingent while holding that another is not, that does not make me more or less contingent than I would be if I held the inverse view.

The categories above provide only half the picture. To complete the taxonomy a second layer is required. Distinctions based on what parts of science are contingent are critical, but we can also, invoking Beatty, draw further distinctions based on what they consider those factors to be contingent upon. Thus, while two people might agree that the methodological components of science are contingent, they might also disagree substantively about the factors upon which methodology is contingent. The factors upon which science, in all its aspects, might be contingent map onto the aspects that can themselves be held contingent: everyday events, sociocultural contexts, methods, interpretations, theories.

4. Summary

I have argued that the debate between contingentists and inevitabilists can be recast as an array of positions that directly oppose one another only over a small range of their total implications. Within the framework provided by Beatty, I have decomposed contingency into seven types, two under unpredictability and five under causal dependence. Each of these latter five might be further decomposed based on the "upon" relation of the contingency in question. These views of contingency can be held alone or in conjunction with others, and each

combination constitutes a distinct position, which carries different assumptions about how science engages with the natural world.

Statements that science is contingent or inevitable are cumbersome when not identifying the area of science on which that property acts and specifying how that property operates within it. Science might be interpretively contingent without being methodologically contingent. It might be both without being theoretically contingent. Many processes play a role in the production of scientific knowledge. Contingency may enter through many doors; it will adopt a different character, with different consequences, when entering through each. The framework I have outlined demonstrates how science can be considered contingent and inevitable in qualitatively different ways and exposes assumptions about the causal structure of the scientific process that would otherwise remain implicit.

References

Beatty, John. 2006. "Replaying Life's Tape." *The Journal of Philosophy* 103:336-362.

Cushing, James. 1994. *Quantum Mechanics: Historical Contingency and the Copenhagen Hegemony*. Chicago: The University of Chicago Press.

Forman, Paul. 1971. "Weimar Culture, Causality, and Quantum Theory: Adaptation by German Physicists and Mathematicians to a Hostile Environment." *Historical Studies in the Physical Sciences* 3:1-115.

Galison, Peter. 1987. *How Experiments End*. Chicago: The University of Chicago Press.

Gould, Steven J. 1989. *Wonderful Life: The Burgess Shale and the Nature of History*. New York: W. W. Norton & Company.

Gould, Steven J., and Lewontin, Richard. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptionist Programme." *Proceedings of the Royal Society of London B* 205:581-598.

Hacking, Ian. 2000. "How Inevitable Are the Results of Successful Science?" *Philosophy of Science* 67:58-71.

Hacking, Ian. 1999. *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Kohler, Robert. 1994. *Lords of the Fly: Drosophila Genetics and the Experimental Life*. Chicago: The University of Chicago Press.

Pickering, Andrew. 1995. *The Mangle of Practice: Time, Agency, and Science*. Chicago: The University of Chicago Press.

Pickering, Andrew. 1984. *Constructing Quarks: A Sociological History of Particle Physics*. Chicago: The University of Chicago Press.

Sankey, Howard. 2008. "Scientific Realism and the Inevitability of Science." *Studies in the History and Philosophy of Science* 39:259-264.

Soler, Léna. 2008a. "Are the Results of Our Science Contingent of Inevitable?" *Studies in the History and Philosophy of Science* 39:221-229.

Soler, Léna. 2008b. "Revealing the Analytical Structure and Some Intrinsic Major Difficulties of the Contingentist/Inevitabilist Issue." *Studies in the History and Philosophy of Science* 39:230-241.

Reconsidering the Argument from Underconsideration[1]

Moti Mizrahi

St. John's University

Abstract

According to the argument from underconsideration, since theory evaluation is comparative, and since scientists do not have good reasons to believe that they are epistemically privileged, it is unlikely that our best theories are true. In this paper, I examine two formulations of this argument, one based on van Fraassen's "bad lot" premise and another based on what Lipton called the "no-privilege" premise. I consider several moves that scientific realists might make in response to these arguments. I then offer a revised argument that is a middle ground between realism and anti-realism, or so I argue.

Keywords

anti-realism, argument from underconsideration, bad lot, epistemic privilege, scientific realism

1. Introduction

The argument from underconsideration is advanced by anti-realists as an argument against scientific realism. According to this argument, it is unlikely that our best scientific theories are true, since theory evaluation is comparative, and since scientists have no good reasons to believe they are selecting from a set of theories that contains a true theory. As Lipton (1993, 89) points out, this argument has two premises. The first is the ranking premise, which states that theory testing yields comparative warrant. As Lipton (1993, 89) puts it: "testing enables scientists to say which of the competing theories they have generated is likeliest to be correct, but does not itself reveal how likely the likeliest theory is."

The second is the no-privilege premise, which states that "scientists have no reason to suppose that the process by which they generate theories for testing makes it likely that a true theory will be among those generated" (Lipton 1993, 89). From these two premises, anti-realists conclude that, "while the best of the generated theories may be true, scientists can never have good reason to believe this" (Lipton 1993, 89). In other words, although they might have good reasons to believe that they have selected the theory that is likeliest to be true from a set of competing theories, scientists have no good reason to believe that any of the competing theories is likely true. The argument from underconsideration is thus aimed against the epistemic thesis of scientific realisms, which is the claim that "Mature and predictively successful scientific theories are well-confirmed and approximately true of the world. So, the entities posited by them, or, at any rate, entities very similar to those posited, inhabit the world" (Psillos 1999, xix).

In what follows, I examine two formulations of this argument, one based on van Fraassen's "bad lot" premise and another based on the "no-privilege" premise. I consider several

[1] This paper has been accepted for presentation at the Philosophy of Science Association meeting in November 2012. Please do not cite without permission.

moves that scientific realists might make in response to these arguments. I then offer a revised argument that is a middle ground between realism and anti-realism, or so I argue.

2. The Bad Lot Premise

According to van Fraassen (1989, 149), scientists may be choosing the best theory of a bad lot. Following Wray's (2010) recent discussion of the argument, van Fraassen's "bad lot" version of the argument can be stated as follows:

(F1) In evaluating theories scientists merely rank the competitors comparatively. [The Ranking Premise]

(F2) There is no reason to suppose that a true theory will be among the theories evaluated. [The Bad Lot Premise]

(F3) Therefore, there is no reason to believe that the theory that is judged to be superior is likely true.

Accordingly, anti-realists claim that there is no reason to suppose that the set of theories to be evaluated contains a true theory. In reply, realists might wonder: why do we need to suppose that? Isn't that what theory testing is all about? Realists might argue that we don't need a reason to think that the set of competing theories contains a true theory before we begin testing. For realists, the testing itself will separate the good theories, if there are any, from the bad ones. If all the theories in the set fail their tests, then it is a bad lot. But if at least one theory passes its tests, then it is not a bad lot after all.

To see why (F2) might seem odd to scientific realists, consider the following analogous argument:

(T1) In evaluating contestants on talent shows, judges merely rank the contestants comparatively.[2]

(T2) There is no reason to suppose that a talented person will be among the contestants evaluated.

(T3) Therefore, there is no reason to believe that the person that is judged to be the winner is likely talented.

Premise (T2) seems rather odd. We do not need to suppose that a talented person is among the contestants. That is what the competition is all about. The competition is supposed to separate the talented from the untalented and weed out the untalented. Like in the case of theory testing, the criterion of selection has to do with success. That is to say, the judges assume that performing excellently on a consistent basis, under the strict conditions of a competition, is a reliable indicator of talent. Again, like in the case of theory testing, if all the contestants fail to perform excellently on a consistent basis throughout the competition, then the lot of contestants is probably a bad one. In any case, it is the competition that will separate the talented from the

[2] I have in mind reality shows in which contestants compete, such as American Idol and Britain Got Talent.

untalented. Similarly, realists would argue, it is experimental and observational testing that will separate the (approximately) true theories from the false ones.

3. The No-Privilege Premise

More recently, Wray (2010) has proposed a revised version of van Fraassen's "bad lot" argument, which was labeled the argument from underconsideration by Lipton (1993). According to Wray (2010, 3), anti-realists argue as follows:

(W1) In evaluating theories scientists merely rank the competitors comparatively. [The Ranking Premise]

(W2) Scientists are not epistemically privileged, that is, they are not especially prone to develop theories that are true with respect to what they say about unobservable entities and processes. [The No-Privilege Premise]

(W3) Hence, we have little reason to believe that the theory that is judged to be superior is likely true.

In response, realists might complain that the no-privilege premise, i.e., (W2), which talks about "epistemic privilege" and scientists being "especially prone," makes it sound as if scientists have a special gift of some sort. But, realists would argue, that is a rather strange way of talking about science. Coming up with good explanations for natural phenomena is a complex human endeavor that involves many factors, having to do with talent, skills, diligence, training, and so on. In addition to the human aspect of theory generation, there is also a methodological aspect involving observation instruments, experimentation techniques, patterns of inference, etc. The no-privilege premise—(W2)—seems to assume that these aspects of theory generation do not change and that scientists never get better at what they do.

To see why (W2) might seem odd to scientific realists, consider the following analogous argument:

(B1) In evaluating desserts, chefs merely rank the competitors comparatively.

(B2) Chefs are not "culinarily privileged," i.e., they are not especially prone to make desserts that are delicious.

(B3) Therefore, we have little reason to believe that the dessert that is judged to be superior is likely delicious.

Premise (B2) seems rather odd. To say that chefs are "culinarily privileged" seems like a strange way of talking about the culinary arts. Chefs get better at making desserts through training and practice. Similarly, realists might argue, scientists get better at developing theories through training and practice. For realists, there is nothing mysterious about "epistemic privilege" going on here. So realists would find (W2) odd for the same reasons that (B2) seems odd.

In reply, anti-realists could appeal to the pessimistic induction. Wray (2010, 6) writes that the "no-privilege thesis […] asks us to acknowledge the similarities between contemporary scientists and their predecessors." He quotes Mary Hesse who argues that the support for the no-

privilege premise comes from an "induction from the history of science." Wray also points out in a footnote that "this is a pessimistic induction of the sort that Laudan (1984) develops." For realists, however, the problem with the pessimistic induction is that it overemphasizes the similarities and underemphasizes the dissimilarities between contemporary theories and their predecessors. Similarly, realists might argue, the problem with Wray's formulation of the argument from underconsideration is that it overemphasizes the similarities and underemphasizes the dissimilarities between contemporary scientists and their predecessors. As Bird (2007, 80) puts it:

> The falsity of earlier theories is the very reason for developing the new ones—with a view to avoiding that falsity. It would be folly to argue that because no man has run 100 m in under 9.5 seconds no man ever will. On the contrary, improvements in times spur on other competitors, encourage improvements in training techniques and so forth, that make a sub 9.5 second 100 m quite a high probability in the near future. The analogy is imperfect, but sufficiently close to cast doubt on Laudan's pessimistic inference. Later scientific theories are not invented independently of the successes and failures of their predecessors. New theories avoid the pitfalls of their falsified predecessors and seek to incorporate their successes.

Likewise, Lipton (2000, 197) argues that we cannot infer "future theories are likely to be false" from "past theories turned out to be false" by induction because of the "Darwinian" evolution of theories. A similar point, realists might argue, applies to scientists as well. Contemporary scientists learn from their predecessors and they seek to avoid their predecessors' mistakes. Furthermore, contemporary scientists have access to instruments and technologies that were not available to their predecessors. For realists, these aspects of scientific change make a difference insofar as the ability of scientists to select theories that are (approximately) true is concerned.

4. Truth vs. Approximate Truth

To this anti-realists might object that the analogous arguments sketched above fail to show that (W2) and (T2) should be rejected, for deliciousness and being talented, which are supposed to be traits analogous to truth, are not analogues to truth at all. Deliciousness and being talented are relative qualities. For example, in the case of deliciousness, whatever cakes we have in a particular lot, we can always imagine being led to consider one of the cakes as delicious, especially if we never tasted a better cake before. But truth is not a relative quality, the objection continues. Propositions are categorically true or false.

In reply, realists might concede that propositions are categorically true or false. However, they might insist that, strictly speaking, only singular propositions can be true or false (Kvanvig 2003, 191), and since theories (whatever they are) are not singular propositions, they cannot be said to be true or false. Accordingly, a theory, expressed as a set of propositions, can have true and/or false propositions as its parts. However, realists might protest, it seems that anti-realists assume that even one false proposition taints a whole theory. For instance, Kitcher points out that the pessimistic induction assumes this kind of implicit holism about theories. As Kitcher (2002, 388) writes:

> We are invited to think of whole theories as the proper objects of knowledge, and thus, because the theory, taken as a whole, turns out to be false, we have the basis for a "pessimistic induction." *It doesn't follow from the fact that a past theory isn't completely* true that every part of that theory is false (emphasis added).

Since only singular propositions can be true or false, and since theories are not singular propositions, it follows that, strictly speaking, whole theories cannot be true or false (Cf. Kitcher 1993, 118).

By way of illustration, consider the following example, which is adapted from Leplin (1997, 133). Suppose that there is a power outage in my house. Upon looking outside my window, I see a utility truck parked nearby and some workers digging in the yard. Since I made a call to the phone company earlier about a problem with my phone line, I infer that telephone repairmen, who have responded to my earlier call, inadvertently cut the power line to my house. Unbeknownst to me, however, it is not telephone repairmen who have cut the power line but cable repairmen whom I had not expected. Now, if we take this "theory," i.e., that there is a power outage in my house because telephone repairmen have inadvertently cut the power line to my house, as a monolithic whole, then it is strictly false. However, this theory involves several claims, some are true and some are false. On the one hand, it is not the case that telephone repairmen working in the backyard have inadvertently cut the power line. On the other hand, it is true that repairmen working in the backyard have inadvertently cut the power line. I may not know the truth, the whole truth, and nothing but the truth about this state of affairs. But I do know some parts about it, and those parts are themselves true.

Consider another example from the history of science. In his An Inquiry into the Causes and Effects of the Variolae Vaccinae (1798), Edward Jenner argues that cowpox originated as grease, a disease common in horses. He claims that it was transmitted to cows when horse handlers helped with milking on occasion. In addition, Jenner (1800, 7) claims not only that cowpox protected against smallpox but also that "what renders the Cow Pox virus so extremely singular, is, that the person who has been thus affected is for ever after secure from the infection of the Small Pox."

Now, if we take the entire Inquiry as Jenner's "theory," then it is strictly false as a whole. He was wrong about grease being the origin of cowpox. He mistakenly took horsepox for grease, and there was no intermediate passage through cows either. Even though he got some things wrong, he was right about others. His hypothesis, properly construed, is correct. While it is not the case that vaccination provides lifelong protection, as Jenner thought, it is the case that repeated vaccination, properly done, contributes to the control of smallpox. Indeed, Jenner paved the way for this knowledge, and the know-how for selection of correct material for vaccination, with his distinction between true and spurious cowpox. Nowadays, pseudocowpox (milker's nodes) is recognized as a type of spurious cowpox (Baxby 1999). According to the World Health Organization, "Publication of the Inquiry and the subsequent promulgation by Jenner of the idea of vaccination with a virus other than variola virus constituted a watershed in the control of smallpox, for which he more than anyone else deserves the credit" (Fenner, et al. 1988, 264).

Another example is Paul Ehrlich's side-chain theory of antibody formation. Ehrlich proposed that harmful compounds can mimic nutrients for which cells express specific receptors.

However, he considered these receptors to be on all cell types. He also did not realize that there are specialized producer cells, such as B lymphocytes. He thought of the entire spectrum of receptors as a single cell because he considered their main task as the uptake of different nutrients. These are parts of Ehrlich's side-chain theory that turned out to be incorrect. It does not follow, however, that the entire theory is wrong. Despite these errors, the theory is based on a correct principle, which is that "specific receptors on cells interact with foreign material in a highly specific way, and this triggers their increased production and release from the cell surface so that they can inactivate foreign material as antibodies" (Kaufmann 2008, 707).

If this is correct, then it seems that we should abandon talk of whole theories as being true or false. Instead, we should talk about theoretical claims as being true or false. Indeed, Wray seems to acknowledge this point. Wray (2008, 323) writes:

> For the sake of clarity, let me call H_1 the Tychonic hypothesis, rather than the Tychonic theory. After all, the Tychonic theory includes an array of other claims (emphasis added).

And, more recently, Wray (2010, 6) writes:

> But our theories, consisting of many theoretical claims, that is, a conjunction of numerous theoretical claims, are most likely false (original emphasis).

If this is correct, then we can distinguish between truth and approximate truth. Articulating a precise notion of approximate truth is beyond the scope of this paper. Nonetheless, on most accounts of approximate truth, this notion is cashed out in terms of a theory being close to the truth. Hence, to say that T is approximately true is to say that T is close to the truth.[3] How do we know that T is close to the truth? Well, realists would argue, we test it. But anti-realists would insist that theory evaluation is comparative. So when we test theories, we compare them. From a set of competing theories, if one theory T passes the tests, then that is a reason to believe that T is closer to that truth than its competitors. If this is correct, then approximate truth, which is a property of theories, is not like truth, which is a property of propositions, insofar as the former is relative, whereas the latter is categorical.

To sum up, then, truth is a property of propositions, since only propositions can be categorically true, whereas approximate truth is a relation between theories, since a theory can be closer to the truth only relative to its competitors. Some might object, however, that theories, expressed as sets of propositions, are simply conjunctions, and conjunctions are categorically truth or false. In reply, I would argue that the truth/approximate truth distinction is analogous to the logical distinction between truth and validity. In logic courses, we teach our students that deductive arguments can be valid or invalid, but not true or false. Even though, in principle, a deductive argument can be expressed as a conditional (i.e., if the premises are true, then the conclusion must be true), which is categorically true or false. In logic, we reserve the terms 'true' and 'false' to premises and conclusions, and the terms 'valid' and 'invalid' to arguments to capture the difference between truth as a property of propositions and validity as a relation between propositions (more specifically, a relation between premises and a conclusion). Similarly, I submit, we should reserve the term 'true' to theoretical claims, which are singular

[3] See, e.g., Leplin (1981), Boyd (1990), Weston (1992), Smith (1998), and Chakravartty (2010).

propositions that can be categorically true or false, and the term 'approximately true' to theories, which is a relation between theories, even though, in principle, theories can be expressed as conjunctions.

5. A Middle-Ground Argument

In Section 3, I have said that realists might find the no-privilege premise—(W2)—in Wray's version of the argument from underconsideration rather odd, since it seems to assume that scientists never get better at theory generation. However, anti-realists might object to that and argue that scientists do get better at theory generation, but they never become good enough such that it is reasonable to believe that their theories are likely true. It seems to me that anti-realists would be correct in arguing that there may not be good reasons to believe that scientists become good enough such that it is reasonable to believe that their theories are likely true. For one thing, the logical space of possible theories is so vast that it seems rather unlikely that scientists would stumble on those competing theories that are closest to the truth. However, I think that anti-realists are wrong in concluding from this that there are no good reasons to believe that certain theories are closer to the truth than others. In this section, then, I will try to carve out a middle ground between realism and anti-realism.

If the aforementioned considerations are correct, then I think it is safe to say that the following claims are true:

(1) Theoretical claims, expressed as singular propositions, can be categorically true or false.

(2) Theories, expressed as sets of propositions, have theoretical claims as their parts.

(3) Scientific theories can be said to be approximately true (i.e., T_1 is closer to the truth than T_2).

(4) Theory evaluation is comparative (i.e., to say that T is approximately true is to say that T is closer to the truth than its competitors).

If these claims are indeed true, as I have argued above, then I think that the following argument can be made, which is a middle ground between scientific realism and anti-realism:

(R1) In evaluating theories, scientists rank the competitors comparatively. [The Ranking Premise].

(R2) If scientists rank competing theories comparatively, then they can only make comparative judgments about competing theories, not absolute judgments (i.e., T_1 is likely true).

(R3) Hence, scientists can only make comparative judgments about competing theories, not absolute judgments (i.e., T_1 is likely true).

(R4) If 'approximate truth' (closeness to the truth) is a relation between theories, then to make comparative judgments about competing theories is to say that a theory is

closer to the truth than its competitors (i.e., T_1 is closer to the truth than T_2, T_3,…, T_n).

(R5) 'Approximate truth' (closeness to the truth) is a relation between theories, not a property of theoretical claims.

(R6) Hence, to make comparative judgments about competing theories is to say that a theory is closer to the truth than its competitors (i.e., T_1 is closer to the truth than T_2, T_3,…, T_n).

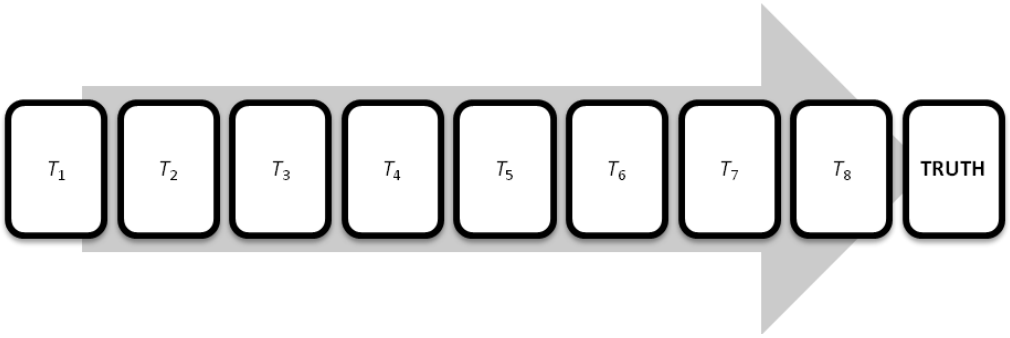
(R7) If the logical space of possible theories is vast, then there are no good reasons to believe that scientists have stumbled upon competing theories that are closest to the truth.

(R8) The logical space of possible theories is vast.

(R9) Therefore, there are no good reasons to believe that scientists have stumbled upon competing theories that are closest to the truth.

The upshot of this argument is that theory evaluation can give us reasons to believe that a theory is approximately true (i.e., that T_1 is closer to the truth than T_2, T_3,…, T_n) but it cannot give us reasons to believe that a theory is closest to the truth (i.e., that T_1 is likely true). For example, if scientists evaluate T_2 and T_3 by observational and experimental testing, they could reasonably make the comparative judgment that T_3 is closer to the truth than T_2 (Figure 1). However, a theory can be closer to the truth relative to its competitors but still be quite far off from the truth. Theory evaluation cannot tell us which theory is closest to the truth, unless we have reasons to believe that the theories we are testing are those that are closest to the truth (i.e., T_7 and T_8 in Figure 1). But, since we do not have reasons to believe that, as anti-realists argue, we cannot reasonably claim that the theories we have tested are closest to the truth (i.e., likely true), although we can reasonably claim that one of them is closer to the truth than its competitors. In other words, theory evaluation can tell us which theory among competing theories is closer to the truth (e.g., that T_3 is closer to the truth than T_2). However, theory evaluation cannot tell us which theory among competing theories is closest to the truth (Figure 1).

Figure 1. T_3 is closer to the truth than T_2 but still quite far off from the truth.



6. Conclusion

In this paper, I examined two formulations of the argument from underconsideration, one based on van Fraassen's "bad lot" premise and another based on what Lipton called the "no-privilege" premise. I considered several moves that scientific realists might make in response to these arguments. I offered a revised argument that I take to be a middle ground between realism and anti-realism, since it adopts the realist thesis that theory evaluation can tell us which theory among competing theories is closer to the truth, and the anti-realist thesis that the lot of competing theories could consist of theories that are far off from the truth, and so theory evaluation cannot tell us which theory is closest to the truth.

References

Baxby, D. 1999. Edward Jenner's Inquiry: A Bicentenary Analysis. Vaccine 17:301-307.
Bird, A. 2007. What Is Scientific Progress? Noûs 41 (1):64-89.
Boyd, R. 1983. The Current Status of the Issue of Scientific Realism. Erkenntnis 19:45-90.
Boyd, R. 1990. Realism, Approximate Truth and Philosophical Method. In Scientific Theories, Minnesota Studies in the Philosophy of Science, edited by C. W. Savage. Minneapolis: University of Minnesota Press.
Chakravartty, A. 2010. Truth and Representation in Science: Two Inspirations from Art. In Beyond Mimesis and Convention: Representation in Art and Science, Boston Studies in the Philosophy of Science, edited by R. Frigg and M. Hunter. Dordrecht: Springer.
Fenner, F., D. A. Henderson, I. Arita, and I. D. Ladnyi. 1988. Smallpox and Its Eradication. History of International Public Health, http://whqlibdoc.who.int/smallpox/9241561106.pdf.
Jenner, E. 1800. An inquiry into the causes and effects of the variolae vaccinae: a disease discovered in some of the western courtiers of England, particularly Gloucestershire, and known by the name of the Cow Pox. 2 ed: Printed for the Author by Sampson Low.
Kaufmann, S. H. E. 2008. Immunology's Foundation: The 100-year Anniversary of the Nobel Prize to Paul Ehrlich and Elie Metchnikoff. Nature Immunology 9:705-712.
Kitcher, P. 1993. The Advancement of Science: Science without Legend, Objectivity without Illusions. New York: Oxford University Press.
Kitcher, P. 2002. Scientific Knowledge. In The Oxford Handbook of Epistemology, edited by P. K. Moser. New York: Oxford University Press.
Kvanvig, J. 2003. The Value of Knowledge and the Pursuit of Understanding. New York: Cambridge University Press.
Laudan, L. 1981. A Confutation of Convergent Realism. Philosophy of Science 48 (1):19-49.
Laudan, L. 1984. Science and Values: The Aims of Science and their Role in Scientific Debate, Pittsburgh Series in Philosophy and History of Science. Berkeley: University of California Press.
Leplin, J. 1981. Truth and Scientific Progress. Studies in History and Philosophy of Science 12:269-291.
Leplin, J. 1997. A Novel Defense of Scientific Realism. New York: Oxford University Press.
Lipton, P. 1993. Is the Best Good Enough? Proceedings of the Aristotelian Society 93:89-104.

Lipton, P. 2000. Tracking Track Records. Proceedings of the Aristotelian Society 74:179-205.
Lyons, T. D. 2006. Scientific Realism and the Stratagema de Divide et Impera. British Journal for the Philosophy of Science 57:537-560.
Psillos, S. 1999. Scientific Realism: How Science Tracks Truth. London: Routledge.
Smith, P. 1998. Approximate Truth and Dynamical Theories. British Journal for the Philosophy of Science 49:253–277.
van Fraassen, B. C. 1989. Laws and Symmetry. Oxford: Clarendon Press.
Weston, T. 1992. Approximate Truth and Scientific Realism. Philosophy of Science 59:53–74.
Wray, K. B. 2008. The Argument from Underconsideration as Grounds for Anti-realism: A Defence. International Studies in the Philosophy of Science 22:317-326.
Wray, K. B. 2010. Epistemic Privilege and the Success of Science. Noûs DOI: 10.1111/j.1468-0068.2010.00793.x.

Bias and Conditioning in Sequential Medical Trials

Cecilia Nardini* Jan Sprenger†

Abstract

Randomized Controlled Trials (RCTs) are currently the gold standard within evidence-based medicine. Usually, they are conducted as *sequential trials* allowing for monitoring for early signs of effectiveness or harm. However, evidence from early stopped trials is often charged with being biased towards implausibly large effects (e.g., Bassler et al. 2010). To our mind, this skeptical attitude is unfounded and caused by the failure to perform appropriate conditioning in the statistical analysis of the evidence. We contend that a shift from unconditional hypothesis tests in the style of Neyman and Pearson to *conditional hypothesis tests* (Berger, Brown and Wolpert 1994) gives a superior appreciation of the obtained evidence and significantly improves the practice of sequential medical trials, while staying firmly rooted in frequentist methodology.

1 Introduction

Randomized Controlled Trials (RCTs) – trials where patients are randomly assigned to a treatment and a control group, while controlling for possible confounders – are currently the gold standard within evidence-based medicine (Worrall 2007). Usually, they are conducted as *sequential trials* allowing for monitoring for early signs of effectiveness or harm.

Monitoring refers to the analysis of data in sequential trials carried out as they accumulate, open to the possibility of stopping the trial before the planned conclusion. By terminating a trial when overwhelming evidence for the effectiveness or harmfulness of a new drug is available we can bound the prohibitive costs of a medical trial and protect in-trial patients

*University of Milan and European Institute of Oncology (IEO), Campus IFOM-IEO, Via Adamello, 16, 20139 Milan, Italy. Email: cecilia.nardini@ieo.eu

†Tilburg Center for Logic and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl

against receiving inferior treatments. Thus, monitoring contributes to meeting ethical and epistemic requirements that clinical investigators are confronted with.

However, monitoring in sequential trials also gives rise to a number of fascinating methodological debates. First, the two grand schools of statistical inference – Bayesian and frequentist inference – are in outright conflict about how to plan and to evaluate a sequential trial. Second, the early termination of sequential trials raises a bulk of concerns: For instance, is it ethically mandatory to stop a trial that indicates the possibility of serious adverse effects, jeopardizing the health of actual patients? Or should the treatment be continued in order to avoid that a successful drug is prematurely rejected, which would deprive future patients of an effective cure?

While we cannot adjudicate these far-reaching questions, we follow Worrall (2008: 418) that "no informed view of the ethical issues [...] can be adopted without first taking an informed view of the evidential-epistemological ones". Thus, we will analyze the statistical methodology of sequential medical trials, focussing on evidence provided by trials *stopped early for benefit*. In the medical literature, such evidence often meets skeptical reactions:

> RCTs stopped early for benefit [...] show implausibly large treatment effects, particularly when the number of events is small. These findings suggest clinicians should view the results of such trials with skepticism. (Montori et al. 2005: 2203)

This standpoint is affirmed by the recent STOPIT-2 metastudy where Bassler et al. (2010: 1187) blame truncated RCTs with "appreciable overestimates of effect". However, we do not share the pessimistic conclusion of these authors. While we believe that some of their criticisms of experimental practice in medicine are valid, we believe that the main issue is not a bias inherent in stopping early for benefit, but the fallacious statistical interpretation of such trials. These misinterpretations are, to our mind, mainly caused by a lack of awareness about issues in statistical methodology that also troubles other disciplines, such as economics and psychology.

Our essay takes the following route. First, we expose the arguments for and against the presence of bias in early stopped trials and explain why this problem is related to principled questions in statistical methodology (Sect. 2). Subsequently, we argue that the real problem is the use of *unconditional error assessments* in sequential trials, rather than the often-invoked divide between Bayesians and frequentists (Sect. 3). Then we show that *conditional*

frequentist tests (e.g., Berger, Brown and Wolpert 1994; Berger 2003) reconcile the need for valid post-experimental appraisal of the evidence with preference for frequentist methods and performance measures in the regulatory framework of medical trials (Sect. 4). Finally, we wrap up our results and sketch how a superior methodological framework can improve the design and practice of sequential trials and eventually lead to better decisions (Sect. 5).

2 Stopping on a random high?

The practice of stopping RCTs early for benefit has been subject to severe epistemological criticism: trials stopped early for benefit show implausibly large treatment effects, relative to what the medical community would be inclined to expect. In a review of 134 trials stopped early for benefit, Montori et al. (2005) point to an inverse correlation between sample size and treatment effect: the smaller the sample size achieved by the trial at the moment of stopping, the larger the estimate it provided for the effect. The more recent study by Bassler et al. (2010) shows that truncated trials report significantly higher effects than trials that were not stopped early.

The danger in stopping a trial for apparent benefit consists in promoting a treatment that is actually less efficacious. For instance, Mueller et al. (2007) report a case of two leukemia treatments where interim analyses suggested a high relative risk reduction (53% and 45%) in a particular chemotherapy regimen. However, that assessment had to be reversed after completion of the trial. In practice, the problem with truncated RCTs is often aggravated by improper reporting: crucial elements of trial design such as sample size, points of the interim analysis, or possible ex-post adjustments of effect estimates are missing in a majority of published trials (Montori et al. 2005).

The aforementioned objections severely threaten the reliability of early stopped trials, as well as their reputation in the medical community. Thus, if investigators wish to stop a trial early, they might do so at the risk of ending up with a result that the medical community does not trust. This situation threatens to nullify the possible advantages of monitoring mentioned in the introduction.

The claim of bias made against trials that stop early is based upon an argument that is known in the medical literature under the name of "stopping on a random high". The argument builds on the consideration that evidence suggestive of a strong treatment effect

can be observed just by chance. Thus, if several interim analyses are performed, sometimes the trial will be stopped for benefit just by chance, exaggerating a small or null effect. It may even be the case that the trend would vanish or even reverse, if the trial were continued, as happened in the leukemia example mentioned earlier.

The validity of this argument has been questioned by several methodologists, especially by those that are familiar with a Bayesian framework. Goodman, Berry and Wittes (2010) argue that the difference observed in the metastudies of Montori et al. (2005) and Bassler et al. (2010) was actually *predictable*: highly efficacious treatments will naturally be more prone to early termination for benefit. Hence, the observed difference in estimated effect size is precisely what we should expect. Comparing early stopped to completed trials amounts, as highlighted by Berry, Carlin and Connor (2010), to selecting the trials to be compared on the basis of their outcome.

Is there a methodologically sound way to account for the worry expressed by the "stopping on random high" intuition? We think that the uneasiness in the medical community is not so much about stopping early, but about trials with implausibly large effects – these effects require, in the words of Mueller et al. (2007), "astute clinicians" to make an appropriate interpretation of the results. In the upcoming section we will argue that this uneasiness is caused by the Achilles' heel of statistical methodology in sequential medical trials: the subscription to unconditional inference procedures.

3 Problems with unconditional inference in sequential medical trials

Sequential medical trials usually control the reliability of a testing procedure from a pre-experimental point of view, by means of Type I and Type II error rates. These error probabilities are extremely important for proper experimental design, and they get a lot of attention from a regulatory point of view. Moreover, frequentist statisticians and philosophers of science have argued that if the sampling plan is violated, the error probabilities cannot be properly controlled and are actually inflated far beyond acceptable (Mayo and Kruse, 2001).

However, adherence to a proper sequential sampling plan is not sufficient to secure a reliable result. As mentioned at the end of the last section, prior knowledge or empirically-

based prior expectations are highly relevant for sound decision-making in the medical arena (cf. Mueller et al. 2007). Yet, at the present state they do not enter the decisions that are ultimately made, except in a methodologically unsatisfactory *ad hoc* way.

In this respect, Bayesian methods have the potential to alleviate the problems with monitoring discussed above. Bayesian reasoners assign a *prior probability distribution* over the values of the parameter of interest (e.g., relative risk reduction). This distribution represents their subjective uncertainty about the true value of the parameter. By means of Bayes' Theorem, this distribution is updated to a *posterior distribution* that synthesizes the observed evidence with the background knowledge.

Goodman (2007) argues that the inclusion of relevant prior information inherent in the Bayesian framework provides a natural way to account for the relevance of contextual knowledge in medical decision-making. From a Bayesian point of view, successful previous studies on a treatment make a positive result for the current trial more expected and thus support the decision to stop early, while on the other hand, negative results of other studies throw a skeptical light on significant observed effects. Thus, unexpected results will be balanced by the prior and lead to a more conservative conclusion than if Bayesian methods had not been used.

In particular, it can be explained that truncated trials provide, ceteris paribus, less *confidence* than trials with a comparable effect size that were completed. The smaller the actual sample, the more will the posterior distribution resemble the prior distribution (for a given effect size). So it appears that the worries of Montori et al. (2005) and Bassler et al. (2010) – overestimation of treatment effect in truncated RCTs – are naturally accounted for.

Despite the advantages just outlined, there are some serious counterarguments to the viability of Bayesianism in clinical trials. A first issue is that some of the philosophical implications of Bayesian inference – such as the evidential, post-experimental irrelevance of experimental design – conflict with the need to carefully plan and conduct sequential medical trials. This is unacceptable to regulatory bodies that are keen to promote proper design of medical trials as a means to ensure the validity of trial results (cf. FDA 2010).

Moyé (2008) has also highlighted a non-sociological point: the problematic specification of a prior belief function.[1] While "objective", non-informative priors (Jeffreys 1961; Bernardo

[1]Similar worries arise regarding the definition of an appropriate loss function required for a Bayesian decision

1979) respond too easily to implausibly large effect sizes, the history of medical trials shows that subjective beliefs about the efficacy of a drug are all too often overturned by surprising findings. The latter problem hampers the use of properly subjective priors and, according to Moyé, it persists even if data from meta-analyses are taken into account.

We consider these worries legitimate and we think they may represent a crucial counter-indication to the use of Bayesian methods in healthcare assessment, even though some of the issues are regulatory rather than epistemological. That said, we believe that the often-cited antagonism between Bayesians and frequentists rests on the false presumption that either of the two is right while the other is wrong. In fact, we suggest to replace that antagonism by the contrast between *conditional* and *unconditional* procedures. By "conditional", we refer to statistical procedures that quantify the conclusiveness of a test result by conditioning on part of the observed data, while "unconditional" refers to the absence of such conditioning.

Arguably, what is most disturbing to the medical community is the fact that, according to current unconditional procedures, a truncated trial has *prima facie* the same reliability as a trial carried to the planned end. This is because Neyman and Pearson's type I and II error rates are unconditional quantities, that is, they are insensitive to whether the data are just at the significance boundary or far beyond it. By contrast, in a conditional perspective, the error associated with a particular conclusion depends on the observed data: the larger the observed difference is, the lower the probability that the null is rejected erroneously.

Practitioners that rely on unconditional inference have an hard time to find informative and reliable *post-data assessments of the evidence*. Often, they report the observed p-value to quantify the conclusiveness of the rejection of the null. However, p-values really combine the worst of all worlds. Since comprehensive and devastating criticisms of using p-values in scientific experiments have been delivered elsewhere (Royall 1997; Goodman 1999), we only mention their most fundamental failures: they neither possess a valid frequency interpretation nor do they provide a useful measure of *confidence* in the null hypothesis.

Moving to *confidence intervals* is often suggested as a way of circumventing the p-value problem (e.g., Cumming and Finch 2005). However, "confidence interval" is a misnomer: a 95% confidence interval merely specifies the set of parameter values that are *consistent* with the observation at the 95% level. This does *not* mean that we should have 95% confidence that

model.

the confidence interval includes the parameter value. In fact, the degree of confidence is just an average coverage rate over intervals from repeated random samples; it is not the coverage probability of the one particular interval that the investigator happens to get. Therefore, it should not come as a surprise that some confidence intervals include the entire sample space, raising the question of what we have actually learned (cf. Seidenfeld 1981).

Finally, we contend that the *unconditional* nature of Neyman-Pearson hypothesis tests is the culprit for their shortcomings. To motivate and to defend this claim, we walk the reader through an example by Cox (1958) and Royall (1997: 74–75).

Suppose that we test $H_0 : \mathcal{N}(0, \sigma^2)$ against $H_1 : \mathcal{N}(1, \sigma^2)$ with known σ^2, and that the toss of a fair coin decides whether we draw N=1 or N=100 i.i.d. observations. It seems natural to apply the most powerful test at the 5% level in either case. However, the probabilistic mixture of the two most powerful tests at the 5% level is *not* the most powerful test in the overall exmperiment. We can do better if we reject H_0 for $x_1 > 1.282$ in the case of N=1, while rejecting H_0 if $\bar{x} > 0.508$ in the case of N=100. Both procedures are tests at the 5% level, but the second, "gerrymandered" test has a greater power (69%) than the mixture of unconditional tests (63%).

Neyman-Pearson methodologists may be inclined to dismiss the second test because not all of its components are tests at the 5% level. However, from an unconditionalist (pre-experimental) viewpoint, only the overall error rates should count. Here, the superior power features speak for the second, gerrymandered test. This problem reveals the tension between the pre-experimental design of unconditional procedures, and the need to efficiently learn from the actual data. Unconditional error rates and confidence intervals do not address that second goal:

> Now if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test [...] is in order. [...] If, however, our objective is to say what we can learn from the data we have, the unconditional test is surely no good. (Cox 1958: 360)

The example can, of course, be easily generalized. It undermines the view that unconditional, pre-experimental error probabilities can qualify the goodness of an inference. In the next section we will see how conditioning on the relevant chunks of information overcomes the problems of unconditional inference and resolves the methodological confusion about

interpreting truncated RCTs, without altering or abandoning the framework of frequentist statistics.

4 Conditional Frequentist Inference

Conditional inference tries to improve upon unconditional procedures by quantifying the degree of confidence that we can have in our conclusions as a function of the observed evidence. More precisely, conditional inference builds on the *strength of the observed evidence*. As we will show in this section, it can be justified from both the Bayesian and the frequentist perspective. The idea comes up for the first time in Cox's (1958) seminal paper, and has been developed later by Kiefer (1977) and Berger (2003), together with various co-authors.

The main idea can be motivated by a very simple example (Kiefer 1977; Berger 2003). Two observations X_1 and X_2 are taken with probability law

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2 \end{cases}$$

If we now construct a confidence interval for θ, then the interval $C_\theta(\cdot, \cdot)$ defined by

$$C_\theta(X_1, X_2) := \begin{cases} X_1 + 1 & \text{if } X_1 = X_2 \\ (X_1 + X_2)/2 & \text{if } X_1 \neq X_2 \end{cases}$$

has an unconditional coverage of 75%. Yet, this does not seem to be a sensible conclusion regarding the *confidence* that the data warrant with respect to the true value of θ. Dependent on whether we observe $|X_1 - X_2| = 0$ or $|X_1 - X_2| = 2$, we are entitled to a statement with (a posteriori) confidence 50% and 100%, respectively. The unconditional coverage of 75% neglects that, after learning the strength of the evidence (that is, the value of $|X_1 - X_2|$), we are in a much better position to assess the confidence which the data grant about our inference. Thus, conditioning on the value of $|X_1 - X_2|$ improves the accuracy of our conclusions.

It is also noteworthy that the probability distribution of $|X_1 - X_2|$ does not depend on the value of θ. That is, $|X_1 - X_2|$ is an *ancillary* statistic with regard to θ. In particular, conditioning on the value of $|X_1 - X_2|$ is quite different from Bayesian conditionalization: where Bayesian change their subjective probability distributions by conditioning on the *entire*

data, conditioning on the value of $|X_1 - X_2|$ just helps to better appreciate the (frequentist) interpretation of the data.

If this idea is applied to hypothesis testing, which is the major issue in medical trials, unconditional error rates are replaced by a conditional error probability. In the following we will outline the basic idea of conditional tests, following Berger, Brown and Wolpert (1994).

Consider, for the purpose of mathematical convenience, the case of testing a point null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$ in some probability model $(\mathcal{X}, \mathcal{B}(\mathcal{X}); \theta \in \Theta)$. Define $f_0(x)$ and $f_1(x)$ as the probability densities of data $x \in \mathcal{X}$ under the hypotheses H_0 and H_1, and let F_0 and F_1 be the corresponding cumulative distribution functions.

$$F_0(x) := P_{H_0}(X \leq x) \qquad\qquad F_1(x) := P_{H_1}(X \leq x)$$

Let the *Bayes factor* $B(x) := f_0(x)/f_1(x)$ be the ratio of the probability density functions, and let

$$\mathcal{X}_s := \{x \in \mathcal{X} | B(x) = s \vee B(x) = F_0^{-1}(1 - F_1(s))\} \tag{1}$$

It is easy to check that $\mathcal{X}_s$ has the same probability density under H_0 and H_1, for all values of s. The intuitive idea is that any $\mathcal{X}_s$ contains two values that have the same strength of evidence under H_0 and H_1. The outcome space is thus partitioned into subsets $\mathcal{X}_s$.

The conditional error probability can now be calculated by conditioning on the particular set $\mathcal{X}_s$ in which the observed data fall. In particular, we can define a *conditional frequentist test* by

$$T^*(X) = \begin{cases} \text{Reject } H_0 & \text{if } B(X) < 1 \\ \text{Accept } H_0 & \text{if } B(X) \geq 1 \end{cases}$$

and for observed $B(x) = s$, we report *conditional error probabilities*

$$\alpha(s) = P_{H_0}(\text{reject } H_0 | X \in \mathcal{X}_s) = \frac{s}{1+s}$$
$$\beta(s) = P_{H_1}(\text{accept } H_0 | X \in \mathcal{X}_s) = \frac{1}{1+s}$$

where the latter equalities have been proven by Berger, Brown and Wolpert (1994, Theorem 1). Clearly, by using the conditional instead of the unconditional error probabilities, we gain a much better appreciation of the chance of a wrong decision, *given the particular data that we have observed*. The higher the Bayes factor, the more confident we can be about an

acceptance of the null, and vice versa. In particular, the classical, unconditional test just detects whether the data are within or outside the rejection region (and leaves the rest to the notorious p-values) whereas the conditional test allows for a fine-grained, properly frequentist discrimination among trials with significant outcomes.

Before moving to the Bayesian interpretation of conditional tests, we would like to briefly discuss a couple of objections that could be made from within the frequentist perspective.

First, it could be argued that T^* makes it far too easy to reject the null ($B(X) < 1$) whereas in medicine, evidence has to be really strong before we are convinced of the efficacy of a new treatment and approve of the drug. To this we simply respond that T^* has been selected because of its simplicity, but it is of course possible to change the rejection region according to contextual requirements.

Second, the use of the Bayes factor may indicate that the conditional test is actually a Bayesian test in frequentist cloths. However, $B(X)$ possesses a frequentist interpretation, too, since it identifies the most powerful frequentist test in the simple vs. simple testing problem.[2]

Third, there may be worries about the *scope* of the above procedure which we have only explained for the easiest possible case of hypothesis testing. However, Berger, Boukai and Wang (1997) have extended conditional tests to simple vs. composite testing problems, and in particular, to the two-sided null hypothesis testing problems that frequently occur in RCTs.

We now explain why T^* is also a valid Bayesian test. Assume that the prior probabilities are balanced: $P(H_0) = P(H_1) = 1/2$. This may be defended as a useful neutrality assumption. Then, the posterior probability of H_0 and H_1 can be written as

$$P(H_0|x) = \left(1 + B(x)^{-1}\right)^{-1} = \frac{B(x)}{1 + B(x)}$$
$$P(H_1|x) = \left(1 + B(x)\right)^{-1} = \frac{1}{1 + B(x)}$$

Thus, we see that the posterior probabilities of H_0 and H_1 correspond to the conditional error probabilities for rejecting H_0 and H_1, respectively. Indeed, the decision to accept H_0 will be wrong whenever H_1 is actually true, that is, with probability $1/1 + B(x)$. Thus, Bayesians and frequentists can conduct the same (conditional) test and obtain the same numerical conclusions. But for the purposed of medical *practice*, philosophical questions about

[2]This is the content of the Neyman-Pearson Lemma. Furthermore Berger (2003) introduced a conditional test that relies on the p-value as the conditioning statistics and yields the same post-data error probabilities as T^*.

the interpretation of probability are clearly secondary as long as there is methodological agreement on procedures and post-experimental data assessment (cf. Berger 2003). In this sense, conditional inference is a genuine reconciliation of Bayesian and frequentist methodology and a real asset for practitioners.

As a last point in indicating the advantages provided by the conditional frequentist framework, we discuss its application to sequential analysis. The proponents of conditional testing have stressed repeatedly that one of the main motivations of conditional inference was the desire to improve upon the practice of sequential testing, particularly in medicine. Here the benchmark is Wald's (1947) famous Sequential Probability Ratio Test, that is

$$T^N(X): \begin{cases} \text{Reject } H_0 \text{ and stop sampling} & \text{if } B(X_1, \ldots, X_N) \leq C^- \\ \text{Accept } H_0 \text{ and stop sampling} & \text{if } B(X_1, \ldots, X_N) \geq C^+ \end{cases}$$

with associated (unconditional) error probabilities

$$\alpha = P_{H_0}(B(X_1, \ldots, X_N) \leq C^-)$$
$$\beta = P_{H_1}(B(X_1, \ldots, X_N) \geq C^+).$$

While these unconditional error probabilities are (i) misleading and (ii) very hard to calculate, Berger, Brown and Wolpert (1994) have suggested a conditional interpretation of this test, choosing C^+ and C^- such that $F_0(C^-) = 1 - F_1(C^+)$, and reporting conditional error probabilities

$$\alpha(B(X_1, \ldots, X_N)) = \frac{B(X_1, \ldots, X_N)}{1 + B(X_1, \ldots, X_N)} \tag{2}$$

$$\beta(B(X_1, \ldots, X_N)) = \frac{1}{1 + B(X_1, \ldots, X_N)}. \tag{3}$$

Thus, the conditional framework can be straightforwardly applied to sequential medical trials, and it has significant advantages. First, the assessment of the error probability depends on the observed data and is thus way more informative than in the unconditional framework. This alleviates the interpretational problem mentioned in Section 2, since conditional error allows medical readers to assess the confidence in the outcome based on the observed data. It seems reasonable to maintain that medical investigators should be more concerned with the actual probability of drawing the wrong inference than with the absolute (unconditional) error rate of the testing procedure.

As a further point, the error probabilities (3) and (4) are independent of the stopping rule, that is the sampling plan determining when the trial is terminated. In a RCT, the stopping rule can never be fully specified, since one cannot cover in advance all eventualities that might happen during a sequential trial. Independence from the stopping rule entails that interpretation of the results and assessment of error are possible even if the stopping rule was misspecified or could not be adhered to due to unforeseen circumstances.

This should not be misunderstood as the claim that pre-data analysis and experimental design are superfluous. Unfortunately, Berger, Brown and Wolpert (1994: 1803) make a claim into that direction, but given the strong emphasis on careful design by methodologists and regulatory bodies (cf. Moyé 2008; FDA 2010), this is unlikely to increase the acceptance of the conditional approach among medical practitioners. We would like to stress that no such claim is required for making a case for the superiority of the conditional frequentist approach. Moreover, since conditional tests can be conducted from both a Bayesian and a frequentist perspective, practitioners do not have to decide for either camp.

There are also interesting implications for the philosophy of statistics: if the "error statisticians" (Mayo 1996) are right that learning from error is indeed a cornerstone of inductive inference, then a move to conditional inference may protect their framework against the objections that we have mentioned in Sect. 3. In particular, there is no need to tie an error-statistical methodology to unconditional inference. However, further developing this line of thought goes beyond the scope of this paper.

5 Conclusions

In this paper we have analyzed the impact of statistical methodology on a substantive ethical and societal question, namely data monitoring in sequential medical trials. In the medical literature, trials stopped early for benefit are often charged with being biased towards implausibly large treatment effects (e.g., Bassler et al. 2010).

We think that this worry is based upon a misinterpretation of sequential trials that is in turn due to shortcomings of standard frequentist procedures. It has been argued (e.g., Goodman 2007) that a Bayesian perspective overcomes this problem: if a trial is stopped early because of an implausibly large effect, blending its result with a (conservative) prior probability distribution naturally mitigates the conclusion. However, as a matter of research

tradition and regulatory requirements – in particular, concerns about individual biases in generating prior distributions –, the Bayesian framework does not provide an easy way out.

In this essay we contend that the real issue is not the contrast between Bayesian and frequentist methodology. Rather, we are concerned about the shortcomings of *unconditional* inference. We have elaborated that while unconditional error probabilities may be helpful in the *design* of an experiment, they do not tell us what we have actually *learned* from the data. We have therefore defended proper conditioning – calculating error probabilities conditional on the strength of the observed evidence – as a way of curing the deficits of unconditional frequentist inference. This approach has a natural application to sequential testing and both a valid Bayesian and a valid frequentist interpretation.

This approach holds considerable promise for the interpretation of early stopped trials in medicine. The possibility of post-data assessments of the probability of an erroneous conclusion represents an invaluable asset for the practitioner and the decision-maker. The results of a medical trial tell much more than the simple acceptance or rejection of a scientific hypothesis: they indicate where evidence is strong and where it is inconclusive, indicating the need for further research. Conditional inference, we believe, can improve the methodology of clinical trials because it allows to take this additional information into account. In conclusion, a clearer view on issues in statistical methodology can help to better appreciate data from sequential medical trials and lead to more efficient and ethically superior decisions in medical research.

References

[1] Bassler, D., Briel, M., Montori, V., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansdell, D., Walter, S., Guyatt, G., N Flynn, D., et al.: Stopping randomized trials early for benefit and estimation of treatment effects. JAMA **303**(12), 1180–1187 (2010)

[2] Berger, J.: Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science **18**(1), 1–12 (2003)

[3] Berger, J., Boukai, B., Wang, Y.: Unified frequentist and Bayesian testing of a precise hypothesis. Statistical Science **12**(3), 133–160 (1997)

[4] Berger, J., Brown, L., Wolpert, R.: A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. The Annals of Statistics **22**(4), 1787–1807 (1994)

[5] Bernardo, J.: Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society. Series B (Methodological) pp. 113–147 (1979)

[6] Berry, S., Carlin, B., Connor, J.: Bias and trials stopped early for benefit. JAMA **304**(2), 156 (2010)

[7] Cox, D.: Some Problems Connected with Statistical Inference. Annals of Mathematical Statistics **29**(2), 357–372 (1958)

[8] Cumming, G., Finch, S.: Inference by eye: Confidence intervals and how to read pictures of data. American Psychologist **60**(2), 170 (2005)

[9] Goodman, S.: Toward evidence-based medical statistics. 1: The P value fallacy. Annals of Internal Medicine **130**(12), 995 (1999)

[10] Goodman, S.: Stopping at nothing? Some dilemmas of data monitoring in clinical trials. Annals of Internal Medicine **146**(12), 882 (2007)

[11] Goodman, S., Berry, D., Wittes, J.: Bias and trials stopped early for benefit. JAMA **304**(2), 157 (2010)

[12] Jeffreys, H.: Theory of Probability. Clarendon Press, Oxford (1961)

[13] Kiefer, J.: Conditional confidence statements and confidence estimators. Journal of the American Statistical Association pp. 789–808 (1977)

[14] Mayo, D.: Error and the growth of experimental knowledge. University of Chicago Press (1996)

[15] Mayo, D., Kruse, M.: Principles of Inference and their Consequences. In: Foundations of Bayesianism. Kluwer Academic Publishers, Netherlands (2001)

[16] Montori, V., Devereaux, P., Adhikari, N., Burns, K., Eggert, C., Briel, M., Lacchetti, C., Leung, T., Darling, E., Bryant, D., et al.: Randomized trials stopped early for benefit: A systematic review. JAMA **294**(17), 2203 (2005)

[17] Moyé, L.A.: Bayesians in clinical trials: Asleep at the switch. Statistics in Medicine **27**, 469–482 (2008)

[18] Mueller, P., Montori, V., Bassler, D., Koenig, B., Guyatt, G.: Ethical issues in stopping randomized trials early because of apparent benefit. Annals of Internal Medicine **146**(12), 878 (2007)

[19] Royall, R.: Statistical Evidence: A Likelihood Paradigm. Chapman & Hall, London (1997)

[20] Seidenfeld, T.: On after-trial properties of best Neyman-Pearson confidence intervals. Philosophy of Science **48**(2), 281–291 (1981)

[21] US Food and Drug Administration: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (2010). Available at `http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm`. Last access 26/01/2012

[22] Wald, A.: Sequential analysis. Wiley, New York (1947)

[23] Worrall, J.: Evidence in Medicine and Evidence-Based Medicine. Philosophy Compass **2**(6), 981–1022 (2007)

[24] Worrall, J.: Evidence and Ethics in Medicine. Perspectives in Biology and Medicine **51**(3), 418–431 (2008)

December 22, 28, 2011
June 20, 2012

The End of the Thermodynamics of Computation: A No Go Result

John D. Norton
Department of History and Philosophy of Science
Center for Philosophy of Science
University of Pittsburgh
Pittsburgh PA 15260
http://www.pitt.edu/~jdnorton
jdnorton@pitt.edu

PSA 2012: Philosophy of Science Association Biennial Meeting
To appear in *Philosophy of Science*.

The thermodynamics of computation assumes that computational processes at the molecular level can be brought arbitrarily close to thermodynamical reversibility; and that thermodynamic entropy creation is unavoidable only in data erasure or the merging of computational paths, in accord with Landauer's principle. The no go result shows that fluctuations preclude completion of thermodynamically reversible processes. Completion can be achieved only by irreversible processes that create thermodynamic entropy in excess of the Landauer limit.

1. Introduction

Electronic computers degrade work to heat and the need for its removal sets a practical limit to their performance. The study of the thermodynamics of computation, surveyed in Bennett (1982), seeks the limits in principle to reduction of this dissipation. Since dissipation reduces with size, the most thermodynamically efficient computers are sought among those that use individual molecules, charges or magnetic dipoles as memory storage devices.

These molecular-scale processes are treated like macroscopic ones in one aspect: they can be brought arbitrarily close to the most efficient, non-dissipative processes, those that are

thermodynamically reversible. Their defining characteristic is that they are at equilibrium at every stage. They are brought slowly from start to finish by the successive nudges of miniscule disequilibria. It is assumed that the dissipative effects of these nudges can be made arbitrarily small by indefinitely extending the time allowed for the process to reach completion.

Some form of dissipation, however, is judged unavoidable. The controlling idea of the thermodynamics of computation is that the creation of thermodynamic entropy and the associated need to pass heat to the environment arise only with logically irreversible operations. These include the erasure of data and the merging of computational paths. The amount of thermodynamic entropy created is quantified by Landauer's principle. It asserts that at least k ln 2 of thermodynamic entropy is created when one bit of data is erased. The result is an elegant account of the bounds to the thermodynamic efficiency of computation. They are independent of the physical implementation, but are set by the logical operations comprising the computation.

Alas, this image of a well-developed science is an illusion. The thermodynamics of computation is an underdeveloped muddle of vague plausibility arguments and misapplications of statistical physics. Earman and Norton (1998, 1999) track the science's history through the Maxwell demon problem and find it rife with circular reasoning and question begging. Norton (2005, 2011) urges that the arguments used to support Landauer's principle are fallacious and have never successfully advanced beyond flawed plausibility arguments. Erasure may reduce the range of possible values for data in a memory. But this reduction is not a compression of the accessible phase space of thermodynamic components that can be associated with a change of thermodynamic entropy. The volume of accessible phase space remains unchanged in erasure. Prior to erasure we may also be unsure as to the data stored and assign probabilities to the possibilities. That sort of probability, however, is not associated with a thermodynamic entropy.

Finally, Norton (2011) describes a "no go" result—that thermodynamically reversible processes at molecular scales are precluded from proceeding to completion by fluctuations. Individual computational steps can only be completed if they are sufficiently far from equilibrium to overcome fluctuations. As a result they create quantities of thermodynamic entropy in excess of those tracked by Landauer's principle. It follows that the lower limit to thermodynamic entropy creation is not set by the logical specification of the computation, but by the details of the particular physical implementation and the number of discrete steps it employs, whatever their function.

This paper will develop the no go result. It is motivated and then stated in the next section. In Section 3, it is illustrated; and in Section 4 a possible loophole is described and closed.

2. The No Go Result

2.1 A Preliminary Form

In a thermodynamically reversible process,[1] all component systems are in perfect equilibrium with one another at all stages. As result, they are impossible processes.[2] Nothing changes. Heat will not spontaneously pass from one body to another if they are at the same temperature. In ordinary thermodynamics, this awkwardness is overcome by introducing a slight disequilibrium. We minutely raise the temperature of the first body and let that minute temperature gradient drive the heat transfer, slowly. Because heat is now passing spontaneously from hot to cold, this is a dissipative process. The thermodynamic entropy created measures the amount of dissipation. For theoretical analyses, this entropy creation can be neglected since it can be made as small as we like by making the driving temperature difference appropriately small. The process will still go forward, but more slowly.

Matters are different when we allow for the molecular constitution of matter. For now the equilibrium of a thermodynamically reversible process is dynamic. If two bodies at the same temperature are in thermal contact, energy will spontaneously pass to and fro between them as energy fluctuations due to random, molecular-scale events. If we are to assure that heat passes

---

[1] Typical erasure processes begin with a thermodynamically irreversible process in which the memory device is thermalized. For example, the wall dividing a two-chamber memory cell is raised so the molecule can access both chambers. The resulting uncontrolled, thermodynamically irreversible expansion creates the k ln 2 of thermodynamic entropy tracked by Landauer's principle. As Norton (2005, Section 3.2) argues, a mistaken tradition misidentifies this thermalization as thermodynamically reversible since the replacing of the partition supposedly returns the original state of "random data."

[2] For an analysis of thermodynamically reversible processes, see Norton (forthcoming, §3).

from the one to the other, we must arrange for a disequilibrium that is sufficiently great to overcome the fluctuations.

Boltzmann's Principle, "S = k ln W," that is, "entropy = k ln probability," measures the dissipation needed. An isolated system is to pass from state 1 with total thermodynamic entropy S_1 to state 2 with total entropy S_2. The inverted principle tells us that, if the system can spontaneously move between the two states, then the probabilities P_1 and P_2 of the two states are related by

$$P_2/P_1 = \exp ((S_2 - S_1)/k) \quad (1)$$

In macroscopic terms, negligible thermodynamic entropy creation is sufficient to drive processes to completion. If $S_2 - S_1 = 10k$, a macroscopically negligible amount, we find $P_2/P_1 = 22{,}026$, so that the final state 2 is strongly favored.

At the molecular level, these amounts of thermodynamic entropy are large. They exceed the entropy change of $k \ln 2 = 0.69k$ tracked by Landauer's principle. They must exceed it, for creation of merely k ln 2 of entropy is insufficient to assure completion of a process. Then $P_2/P_1 = \exp (k \ln 2/2) = 2$. The process is only twice as likely to be in its final state 2 as in its initial state 1. This is a fatal result for the thermodynamics of computation. If we have any computing process with multiple steps operating at molecular scales, we must create thermodynamic entropy in each step if the process is to go forward, quite aside from any issues of logical irreversibility.

2.2 The Main Result

Boltzmann's Principle in the form (1) applies to isolated systems. In the thermodynamics of computation, the computing systems are treated as open systems, in equilibrium with a heat bath at the ambient temperature T. The main result arises when we adapt these considerations to such systems.

A computer is a system consisting of many interacting components, including memory cells, systems that read and write to the memory cells and other control components to implement the computer's program. At any moment, the combined system is in thermal equilibrium with the environment at temperature T. Hence, the system is canonically distributed over its phase space, according to the probability density

$$p(\mathbf{x}, \boldsymbol{\pi}) = \exp(-E(\mathbf{x}, \boldsymbol{\pi})/kT)/ Z$$

where Z is the normalizing partition function and $\mathbf{x}$ and $\boldsymbol{\pi}$ are multi-component generalized configuration and momentum coordinates.

Each computational step is carried out by a thermodynamically reversible process, whose stages are parameterized by λ. Fluctuations will carry the system spontaneously from one stage to another. As a result, the system is probabilistically distributed over the different stages. The probabilities are computed by Einstein's methods, as adapted by Tolman (1938, pp. 637-38), and conform to the probability density

$$p(\lambda) = \text{constant}. Z(\lambda) \tag{2}$$

where $Z(\lambda)$ is given by

$$Z(\lambda) = \int_{\lambda} \exp(-E(\mathbf{x}, \boldsymbol{\pi})/kT)\, d\mathbf{x}d\boldsymbol{\pi}$$

This last integral extends over the volume of phase space accessible to the system when the process is at stage λ.

In the Einstein-Tolman analysis, each of these stages is given a thermodynamic description as if it were an equilibrium state, even though it may have arisen through a fluctuation. The canonically distributed system at stage λ is assigned a canonical free energy

$$F(\lambda) = -kT \ln Z(\lambda) \tag{3}$$

treating $Z(\lambda)$ as a partitition function, where the free energy is defined as

$$F(\lambda) = E(\lambda) - TS(\lambda)$$

Here $E(\lambda)$ and $S(\lambda)$ are the mean energy and the thermodynamic entropy assigned to the system in stage λ. It now follows from (2) and (3) that

$$p(\lambda) = \text{constant}. \exp(-F(\lambda)/kT)$$

and that the probability densities for the system fluctuating between stages λ_1 and λ_2 satisfy

$$p(\lambda_2)/ p(\lambda_1) = \exp(-(F(\lambda_2) - F(\lambda_1))/kT) \tag{4}$$

The process is thermodynamically reversible. Hence it is in equilibrium at every stage. Equilibrium requires the vanishing of the generalized thermodynamic force $X(\lambda)$ acting on the system:[3]

$$X(\lambda) = - \partial/\partial\lambda\big|_T F(\lambda) = 0$$

Integrating over λ, we find that the free energy $F(\lambda)$ is constant over the stages of the process:

$$F(\lambda) = \text{constant} \qquad F(\lambda_1) = F(\lambda_2) \tag{5}$$

From (4), we have that

$$p(\lambda) = \text{constant} \qquad p(\lambda_1) = p(\lambda_2) \tag{6}$$

This last result (6) is the no go result. It precludes thermodynamically reversible processes proceeding as we expect.

Our default expectation is that these processes are in a quiescent equilibrium at every stage λ, perhaps with a slight disturbance due to fluctuations. We expect to bring the process from its initial to its final stage by minute disequilibrium nudges that advance the process arbitrarily slowly in the tiniest of steps. What (6) tells us is that fluctuations obliterate the quiescent equilibrium. If the system is in one stage λ at some moment, it is equally likely to be found at the next moment at any other stage. If we set up the process in its initial stage, it is as likely to leap by a fluctuation to the final stage as it is to stay where it is. If the process has arrived at the final stage, it is as likely to be flung by a fluctuation back to its initial stage, as it is to stay where it is. In a slogan, fluctuations obliterate thermodynamically reversible processes.

Fluctuations are temperature sensitive. Hence we might expect the confounding effects of fluctuations to be calmed and controlled by cooling the processes, perhaps even close to absolute

---

[3] At equilibrium, the total entropy S_{tot} of the system S_{sys} and the environment S_{env} is stationary. Writing $d = \partial/\partial\lambda\big|_T$, that amounts to $0 = dS_{tot} = dS_{sys} + dS_{env}$. By supposition, the computer system exchanges no work with the environment, but only heat in a thermodynamically reversible process. Hence $dS_{env} = dE_{env}/T = - dE_{sys}/T$, where the last equality follows from conservation of energy: $dE_{env} + dE_{sys} = 0$. Combining, we have $0 = dS_{sys} - dE_{sys}/T$. Hence the condition for equilibrium is $0 = d(E_{sys} - TS_{sys}) = -X_{sys}$.

zero. A review of the calculation above shows that the no go result (6) obtains no matter what the temperature, even if it close to absolute zero.[4]

2.2 What It Takes to Beat Fluctuations

If fluctuations obliterate thermodynamically reversible processes, how is it possible for these processes to figure in thermodynamic analysis at all? The answer is that the disequilibrium required to overcome fluctuations is negligible macroscopically. While the no go result applies to macroscopic systems, it is overcome by disequilibria too small to trouble us. However, at the molecular scale explored by the thermodynamics of computation, the situation is reversed. There, the disequilibria needed to overcome fluctuations dominate. Most importantly, it requires thermodynamic entropy creation in amounts that well exceed those tracked by Landauer's principle.

A few computations illustrate this answer. Relation (4) tells us that we can probabilistically favor the end stage λ_2 over the initial stage λ_1 if the end stage free energy $F(\lambda_2)$ is smaller than the initial stage free energy $F(\lambda_1)$. A decrease of 3kT is sufficient for a modest favoring in the ratio of 20:1, for then

$$p(\lambda_2)/p(\lambda_1) = \exp(-(-3kT)/kT) = \exp(3) = 20$$

The dissipation associated with the reduction in free energy $F(\lambda_2) - F(\lambda_1) = -3kT$ is a minimum increase in the thermodynamic entropy of[5]

---

[4] Temperature does affect the free energy needed to override the fluctuations. We see below that a probabilistic favoring of 20:1 is achieved by a free energy reduction of 3kT. This reduction diminishes as T decreases. However the thermodynamic entropy created remains at least 3k, independent of the temperature.

[5] To see this, use F=E-TS to rewrite $F(\lambda_2) - F(\lambda_1) = -3kT$ as

$$S(\lambda_2) - S(\lambda_1) - (E(\lambda_2) - E(\lambda_1))/T = 3k$$

We have $\Delta S_{sys} = S(\lambda_2) - S(\lambda_1)$. By conservation of energy, $-(E(\lambda_2) - E(\lambda_1))$ is the energy gained by the environment. By supposition, this energy is passed by heat transfer only. In the least dissipative case of a thermodynamically reversible heat transfer that corresponds to the minimum increase of entropy $\Delta S_{env} = -(E(\lambda_2) - E(\lambda_1))/T$.

$$\Delta S_{tot} = \Delta S_{sys} + \Delta S_{env} = 3k$$

where the change Δ is applied to the entropy of the universe as a whole S_{tot}, which is the sum of the system entropy S_{sys} and the environment entropy S_{env}. Even though this modest probabilistic favoring by no means assures completion of the process, the entropy creation of at least 3k is many times greater than the k ln 2 = 0.69k of entropy tracked by Landauer's principle in a single bit erasure.

Since the ratio of probability densities grows exponentially with free energy differences in (4), further creation of thermodynamic entropy can bring probability density ratios that strongly favor completion of the process. For example, if we increase the free energy difference to 25kT, then the end stage is strongly favored, for

$$p(\lambda_2)/p(\lambda_1) = \exp(-(-25kT)/kT) = \exp(25) = 7.2 \times 10^{10}.$$

In macroscopic terms, however, 25kT of free energy is negligible. This quantity, 25kT, is the mean thermal energy of ten diatomic molecules, such as ten oxygen molecules. Hence, there is no obstacle to introducing a slight disequilibrium in a macroscopic system in order to nudge a thermodynamically reversible process to completion.

3. Illustrations of the No Go Result for a One-Molecule Gas

This no go result applies to all thermodynamically reversible processes in systems in thermal equilibrium with their environment. However its derivation and its statement as (6) is remote from its implementation in specific systems. It is helpful to illustrate how fluctuations obliterate a simple process described in the thermodynamics of computation, the thermodynamically reversible, isothermal expansion and compression of a one-molecule gas. The analysis of the last section provides the precise computation. Here I give simpler estimates of the disturbing effects of fluctuations.

3.1 Reversible, Isothermal Expansion and Compression

A monatomic one-molecule gas is confined to a vertically oriented cylinder and the gas pressure is contained by the weight of the piston. The process intended is a thermodynamically reversible, isothermal expansion or compression of the gas. Our expectation is that this process will proceed indefinitely slowly, with the weight of the piston maintained just minutely away

from the equilibrium weight so that the expansion or compression is only just favored. As the piston is raised in an expansion, it draws work energy from the one-molecule gas; and this energy is restored to the one-molecule gas as heat from the environment. The gas exerts a pressure P=kT/V, for V the volume of the gas. Thus the work extracted in a doubling of the volume and thus also the heat passed to the gas is given by $\int_V^{2V} kT/V'dV' = \text{kT ln 2}$. The thermodynamic entropy change in the gas is the familiar k ln 2.

That is our expectation. It is confounded by fluctuations. Consider the piston first. It is a thermal system that is Boltzmann distributed over its height $h \geq 0$ above the piston floor according to

$$p(h) = (Mg/kT) \exp (-Mgh/kT)$$

where M is the piston mass. The mean of this distribution is kT/Mg and its standard deviation is also kT/Mg.

This latter number measures the extent of thermal fluctuations in the height of the piston. For a macroscopic piston, M will be very much larger than kT/g and the extent of fluctuations in height will be negligible. However in this case of a one-molecule gas, the piston must be very light if it is to be suspended at equilibrium by the pressure of the one-molecule gas. Hence its M is small and the fluctuations in height will be great. They can be estimated quantitatively as follows. The weight of the piston is Mg. The mean force exerted by the gas pressure is (kT/V).A = kT/h, where A is the area of the piston and h its height above the base of the cylinder, so that V = Ah. Setting these two forces equal as the condition for equilibrium, we recover the equilibrium height as[6]

$$h_{eq} = kT/Mg$$

Remarkably, this quantity h_{eq} is just the same as the mean height and standard deviation of the above distribution, both of which are also given by kT/Mg.

---

[6] Hence the mean energy of height is $Mgh_{eq} = kT$. While this energy is associated with a single degree of freedom of the moving piston, it differs from the familiar equipartition mean energy per degree of freedom (1/2)kT, because the relevant term of the piston's Hamiltonian, Mgh, is linear in h and not quadratic, as the equipartition theorem assumes.

This extraordinary result can be expressed more picturesquely as follows. If we set up the piston so that its weight perfectly balances the mean pressure force of the one-molecule gas, it will not remain at the equilibrium height, but will fluctuate immediately through the entire volume of the gas. It will perhaps be suddenly flung skyward by a collision with molecule; and it may then fall precipitously between collisions. The intended process of a gentle, indefinitely slow expansion or contraction is lost completely behind the wild gyrations of the piston over the full volume of the one-molecule gas.

Similar results hold for heat transfer between the one-molecule gas and its environment. Since it is monatomic, the Boltzmann distribution of the gas energy E is

$$p(E) = 2(E/\pi)^{1/2} (kT)^{-3/2} \exp(-E/kT)$$

The mean of this distribution is the familiar equipartition energy (3/2) kT and the standard deviation is $(3/2)^{1/2}$ kT = 1.225 kT.[7] Hence, simply by virtue of its contact with the environment at temperature T, the one-molecule gas energy will be swinging wildly through a range comparable in size to the total mean energy of the gas.

We had expected that we would track a quantity of heat kT ln 2 = 0.69 kT while the piston slowly and gently moves to halve or double the volume of the gas. What we find is that the piston is wildly and randomly flung to and fro through the entire volume of the gas, while the gas energy fluctuates similarly wildly over a range greater than the 0.69 kT of heat transfer we track. We had expected a process that proceeds calmly at arbitrarily slow speed from start to finish. Instead we find a chaos of wild gyrations with no discernible start or finish.

This is a rough analysis. To maintain the equilibrium of a thermodynamically reversible process would require that the weight Mg be adjusted as the volume V changes since the gas pressure will vary inversely with volume. Norton (2011, Section 7.5) replaces the uniform force field of gravity with another force field that varies with height in precisely the way needed to maintain mean quantities at equilibrium.

---

[7] This and the earlier energy standard deviation can be computed most rapidly from Einstein's energy fluctuation theorem, which identifies the variance of the energy with $kT^2\, d\langle E\rangle/dT$, where $\langle E\rangle$ is the mean energy. For the piston, $\langle E\rangle = kT$, so the variance is $(kT)^2 = (Mgh_{eq})^2$. For the monatomic gas, $\langle E\rangle = (3/2)kT$, so the variance is $(3/2)(kT)^2$. The standard deviation is the square root of the variance.

3.2 Generality

A one-molecule gas confined in a cylinder by a piston is fanciful and cannot be realized practically. It is, however, one of the most discussed examples in the thermodynamics of computation because it is easy to visualize. Its statistical and thermodynamic properties mimic those of more realistic systems with few degrees of freedom. We may model a memory device as a two-chambered cell with a single molecule trapped in one part. A more realistic implementation of the memory device is a single electric charge trapped by a potential well in a solid state medium; or a magnetic dipole aligned into a specific orientation by a magnetic field.

The thermodynamic operations carried out on the one-molecule gas have analogs in the more realistic implementations. Mechanical variables such as volume and pressure are replaced by electric and magnetic correlates. The general results remain the same. If we halve the range of possible states of a memory device, we reduce its thermodynamic entropy by k ln 2, just as we do when we halve the volume of a one-molecule gas. The large fluctuations exhibited by the one-molecule gas derive from its small number of degrees of freedom. Correspondingly, the more realistic implementations will exhibit similarly large fluctuations.

The two processes investigated were heating/cooling and expansion/contraction of the gas. These are instances of the two processes that appear in all thermodynamically reversible processes: heat transfer and exchange of generalized work energy. As a result, the analysis here has a quite broad scope. Consider thermodynamically reversible measurement, in which one device reads the state of another. For example, a magnetic dipole reads the state of a second dipole when the two slowly approach and align in a process that maintains equilibrium throughout. This detection or measurement process is a reversible compression of the phase space of the reader dipole and is thermodynamically analogous to compression of a one-molecule gas. As a result, this measurement process will be fatally disrupted by fluctuations. While a standard claim of the thermodynamics literature is that these measurements can be performed without dissipation, the no go result shows that dissipation is required if the fluctuations are to be overcome and the process driven to a correct reading.

4. A Loophole?

Each computation consists of many steps. Dissipation, significant at the molecular level, is required by the no go result to bring each of these steps to completion. Bennett (1973, 1982) proposes an ingenious loophole for computations with very many steps. The very many thermodynamically reversible steps are chained together to form one large thermodynamically reversible process. The computer's state wanders back and forth through the various stages in a generalization of Brownian motion. The no go result affirms that the state will be uniformly distributed over all the stages of the computation. Bennett now makes the step to the final state highly dissipative, so that it can be favored with arbitrarily high probability. Hence the computation will eventually terminate in this final state with high probability. The thermodynamic entropy created in this final, irreversible step may be large. However, if there are very many steps combined into the overall computation, the entropy created per step can be quite small.

Whether this loophole can succeed depends on whether the many steps of a computation can be chained together in such a way that achieving the final state also assures that all the computer's components are in the intended final states. The danger point is when the computer completes one step and initiates the next. The initiation of the second step must arise only when the first step is completed and the state of the computer conforms to what the logical specification of the program requires for that first step. We need to be assured that the disrupting effects of fluctuations will not trigger the second step before these conditions are met.

In an attempt to assure this, Bennett (1982) describes a Brownian clockwork computer, a mechanical implementation of a Turing machine. Its parts are mechanically interlocked so that when the tape manipulator head reaches its final state, each of the cells of the tape are in the final states intended by the logical specification of the computation.

Bennett's description of the device is detailed with vivid line drawings. However it is incomplete in the one aspect that matters most. The statistical mechanical properties of the individual components are poorly represented. Here is the easiest way to see that they are omitted: the machine is sufficiently powerful that we could set it up with a large tape carrying

"random" data of 0s and 1s and then run an erasure program that resets all the cells to zero.[8] On Bennett's view, there must be an associated creation of thermodynamic entropy of at least k ln 2 per bit erased and the passing of kT ln 2 of heat per bit erased to the environment. Yet their creation is nowhere apparent in the operation of the machine.[9]

The narrative that describes the machine's operation depends on our imagining processes that are unproblematic if implemented by macroscopic bodies. For example, the branching of the program's execution arises when the path of the manipulator is obstructed by a knob whose position encodes the data recorded in the tape cell. Our macroscopic intuitions preclude the manipulator ever proceeding with a misread of the data. These same processes may fail if we attempt to implement them in a thermodynamically reversible manner at the molecular level. For that means that all interactions must be at equilibrium. The components at issue, such as a single molecule or a molecular-scale dipole, exert very weak forces on average and these forces are confounded by fluctuations comparable in size to the average. Another component interacting with them can only apply correspondingly weak forces, else the requirement of equilibrium of thermodynamic reversibility would be violated. Once again our intended average behavior would be immersed in wild fluctuations. The resulting interaction would be very different from a macroscopically pictured manipulator thumping into macroscopic knob and being definitively obstructed by it.

The following indicates how adding these thermal complications would compromise the operation of the clockwork computer. The obstruction of the manipulator head by the data knob is equivalent to the reading by a detector of the state of a data cell. The manipulator in effect reads the state of the data cell and records the reading by implementing one of several possible computational paths. Bennett (1982, pp. 307-308; 1987, p.14) has described two schemes in which a reader detects the position of a single component memory device in a reversible thermodynamic process. The molecular implementation is quite fragile in comparison with its robust macroscopic counterpart and fails precisely because the analysis of both schemes neglects

---

[8] The program reads a cell and rewrites its contents to 0, if the cell has a 1. If the cell has a 0, it moves one cell to the right and repeats.

[9] Or one could assume that the physical description is complete so that the machine can erase the tape without thermodynamic entropy creation. That contradicts Landauer's principle.

how fluctuations confound the intended behavior of thermodynamically reversible processes at the molecular scale. Norton (2011, §7.3) describes how both detection schemes fail. For the case of binary data, they are as likely as not to terminate with the detector reading the right as the wrong result.

We have every reason to expect that these problems would appear were the clockwork, Brownian computer somehow implemented with molecular scale storage devices and operated by thermodynamically reversible processes. We have no assurance that any step would proceed according to its logical specification. If the reading of data in a cell is implemented as Bennett describes, they would likely as not return the wrong result. When the manipulator is eventually trapped probabilistically in its final state, we should expect the tape to be left in a state of chaos that does not reflect the results intended by the logical specification of the program.

In short, the loophole fails. It is a conjecture, motivated by macroscopic intuitions that do not apply at molecular scales.

References

Bennett, Charles. 1973. "Logical Reversibility of Computation." *IBM Journal of Research and Development,* **17**, 525-32.

Bennett, Charles. 1982. "The Thermodynamics of Computation—A Review." *International Journal of Theoretical Physics*, **21**, 905-40; reprinted in Leff and Rex, 2003, Ch. 7.1.

Bennett, Charles, H. 1987. "Demons, Engines and the Second Law." *Scientific American*, **257**(5), 108-116.

Earman, John and Norton, John D. 1998. "Exorcist XIV: The Wrath of Maxwell's Demon." Part I "From Maxwell to Szilard." *Studies in the History and Philosophy of Modern Physics*, **29**(1998), 435-471.

Earman, John and Norton, John D. 1999. "Exorcist XIV: The Wrath of Maxwell's Demon." Part II: "From Szilard to Landauer and Beyond," *Studies in the History and Philosophy of Modern Physics*, **30**(1999), 1-40.

Leff, Harvey S. and Rex, Andrew, eds. 2003. *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing*. Bristol and Philadelphia: Institute of Physics Publishing.

Norton, John D. 2005. “Eaters of the lotus: Landauer's principle and the return of Maxwell's demon.” *Studies in the History and Philosophy of Modern Physics*, **36**, 375–411.

Norton, John D. 2011. “Waiting for Landauer.” *Studies in History and Philosophy of Modern Physics,* **42**, 184–198.

Norton, John D. Forthcoming. "Infinite Idealizations," Prepared for *Vienna Circle Institute Yearbook*. Springer: Dordrecht-Heidelberg-London-New York.

Tolman, Richard C. .1938. *The Principles of Statistical Mechanics*. London: Oxford University Press.

Abstract

This paper provides a new context for an established metaphysical debate regarding the problem of persistence. I contend that perdurance (the claim that objects persist by having temporal parts) can be precisely formulated in quantum mechanics due to an analogy with spatial parts, which I claim correspond to the decomposition of the quantum state provided by a localization scheme. However, I present a 'no-go' result that rules out the existence of an analogous temporal localization scheme, and so argue that quantum objects cannot be said to perdure. I conclude by surveying the remaining metaphysical options.

Do Quantum Objects Have Temporal Parts?

1st March, 2012.

1. Introduction This paper provides a new context for an established metaphysical debate regarding the *problem of persistence*. Namely, how can it be said that one and the same physical object persists through time while changing over time? I contend that a popular view about persistence which maintains that objects persist by *perduring* – that is, by having temporal parts – receives a particularly neat formulation in quantum mechanics due to the existence of a formal analogy between time and space. However, on closer inspection this analogy fails due to a 'no-go' result which demonstrates that quantum systems can't be said to have temporal parts in the same way that they have spatial parts. Therefore, if quantum mechanics describes persisting physical objects, then those objects cannot be said to perdure.

This argument serves two aims. The first is to continue the recent tradition of addressing the problem of persistence in the context of specific physical theories: Balashov (2010) considers special relativity; Butterfield (2005, 2006) considers classical mechanics. The second aim is to provide a novel interpretation of the no-go result mentioned above, which is well-known in the quantum foundations literature but rarely discussed by philosophers of physics. The result is often phrased like this: There exists no time observable canonically conjugate to the Hamiltonian. This fact was first observed by Pauli in 1933, and there are

various proofs which arrive at this conclusion.[1] I claim this result is best understood not as an argument against the existence of time (Halvorson 2010) but rather as an argument that quantum systems do not have (proper) temporal parts.[2]

1.1. Argument Outline The argument takes the form of a *modus tollens* which I give a sketch of here, leaving technical details for later sections. I begin with a characterization of *perdurance* as the thesis that objects persist through time just as they stretch through space.

perdure The part-whole relation for persisting objects applied to time works just like the part-whole relation with respect to space. That is, persisting objects have proper temporal parts associated with an arbitrary division of the times over which the object persists.

I then argue that ordinary quantum mechanics describes persisting objects. This claim requires (at least) a robust scientific realism about quantum mechanics.

quantum A quantum object (an isolated system described by a ray in Hilbert space undergoing unitary evolution) is a persisting object.

Taken together, **perdure** and **quantum** imply that quantum objects have temporal parts. Call this view *quantum perdurantism*. I further claim that if quantum mechanics does describe

[1]It was recently observed that Pauli's proof admits a significant class of counterexamples (Galapon 2002). The result I will give is instead related to the proofs of Srinivas and Vijayalakshmi (1981); Halvorson (2010).

[2]This is not to say that there are not other valid interpretations. For example, Unruh and Wald (1989) provide an argument against the existence of an ideal quantum clock.

persisting objects, then it also provides a legitimate account of the spatial parts of such an object.

spatial The part-whole relation in space for quantum objects is given by the *spectral decomposition* of the Hilbert space into orthogonal subspaces, provided by the position operator $\hat{Q}$. This decomposition is unique, and is known as a localization scheme. A quantum object has spatial parts *iff* there exists a localization scheme.

Since **perdure** asserts that the relation of parthood applied to time is just like the relation of spatial parthood, it follows that a quantum perdurantist is committed the existence of a *temporal* localization scheme which operates in an analogous way to **spatial**. That is, **perdure**, **quantum** and **spatial** jointly entail the following conditional statement.

temporal If quantum perdurantism is true then every persisting quantum object has a unique decomposition into temporal parts provided by a temporal localization scheme.

Unfortunately for the perdurantist, the consequent is demonstrably false: The *spectral condition* states that the Hamiltonian of every system has a spectrum bounded from below – roughly, every system has a state of lowest energy – and entails that no quantum objects possess a temporal localization scheme. Therefore quantum perdurantism is false; quantum objects have no (proper) temporal parts.

This leaves two possibilities: Either they have no temporal parts (*endurantism*), or one temporal part (*temporal holism*). I argue that, although there is little to choose between these rival views in the context of non-relativistic quantum mechanics, considerations from relativity favor temporal holism.

2. The Metaphysics of Persisting Objects How does a material object persist *through* time while changing *with* time? There are essentially two schools of thought: Either a persisting object has no temporal parts, is self-identical at every moment it exists, and its spatial properties change with time (*endurantism*), or a persisting object necessarily has temporal parts which have differing spatial properties (*perdurantism*). Another common way of phrasing the distinction is as a conflict between *three-dimensionalism* and *four-dimensionalism*: If an object endures then it exists in three dimensions (since it has no temporal width); if it perdures then, having temporal width, it exists in four dimensions.[3]

I will follow Lewis (1986) in using the term perdurance for the latter possibility, which I take to be a thesis about the existence of temporal parts.

> Something perdures iff it persists by having different temporal parts, or stages, though no one part of it is wholly present at more that one time; whereas it endures iff it persists by being wholly present at more than one time. Perdurance corresponds to the way a road persists through space; part of it is here and part of it is there, and no part is wholly present at two different places. (Lewis 1986, 202)

So for perdurantism to be true, it must be the case that persisting objects be amenable to decomposition into (proper) temporal parts. It has been complained that the notion of *being wholly present* is problematic (*e.g.* (Sider 1997; McCall and Lowe 2003)) but I will argue that, within quantum mechanics, it can be given a precise meaning due to a formal analogy with *being wholly located*.

[3]Among four-dimensionalists, there is a further dispute about reference: When we speak of an object *e.g.* "the table" do we refer to a particular instantaneous *stage*, *i.e. the table at time* t, (Sider 1997; Hawley 2004) or the entire temporally extended object?

I take the motivation for perdurantism to be a strong analogy between time and space. Sider expresses this idea as follows:

> As I see it, the heart of four-dimensionalism [perdurantism] is the claim that the part-whole relation behaves with respect to time analogously to how it behaves with respect to space ... Applied to time, the idea is that for any way of dividing up the lifetime of an object into separate intervals of time, there is a corresponding way of dividing the object into temporal parts that are confined to those intervals of time. (Sider 1997, 204)

I will argue that the appropriate part-whole relation for spatial parts in quantum mechanics is provided by a localization scheme (in Section 4), which commits the perdurantist to a thesis about temporal *localizability* in quantum mechanics (in Section 5).

3. Persisting Objects in Quantum Mechanics Quantum mechanics provides our best theory of matter, and its empirical predictions have been startlingly accurate. That much is uncontroversial. On the other hand, any attempt to assert exactly *why* it has proved so successful, or precisely what it tells us about the nature of material objects involves taking sides on disputes regarding its interpretation that have lasted over 80 years and show little sign of abating. Therefore, I will proceed by specifying under what conditions one would be committed to regarding the quantum state as describing a persisting material object. Nonetheless, I take it that *prima facie* a realist metaphysician who takes chairs (composed of complex collections of organic molecules) to be persisting objects would be compelled to similarly regard, say, a molecule of Buckminsterfullerene (C_{60}) composed of sixty atoms of carbon, and recently shown to display distinctly quantum behavior (Nairz et al. 2003).

First, some details about the formalism of ordinary (non-relativistic) quantum mechanics. As our concern is with spatio-temporal properties, we will consider systems with no internal degrees of freedom (*i.e.* spinless particles). Therefore, the *state space* of the theory is provided by the space of square integrable functions defined over all of space, that is, infinite-dimensional (separable) complex Hilbert space $\mathcal{H} = L^2(\mathbb{R})$ (for simplicity we will consider only one spatial dimension). The *pure* states $|\psi\rangle$ are in one-to-one correspondence with the one-dimensional subspaces of $\mathcal{H}$ or, equivalently, the set of normalized vectors that individually span those subspaces. Since $\mathcal{H}$ is a vector space, linear combinations of pure states are also pure states (this is known as the *superposition principle*). In what follows I will only consider pure states.

The first interpretative posit I require is *realism*, the claim that real physical systems are authentically described by quantum mechanical states. The next posit I require is *completeness*, the claim that a pure state provides a complete description of an individual quantum system which leaves nothing out (*i.e.* no hidden variables). So far we would be justified in claiming that the quantum state describes a physical object. But what about *persisting* objects?

We require some facts about quantum dynamics. In the Schrödinger picture, the history of a system is given by a series of (pure) states $|\psi(t)\rangle$, where $t \in \mathbb{R}$. Once the state $|\psi(0)\rangle$ at a time $t = 0$ is given, the entire history is determined according to the time-dependent Schrödinger equation in terms of a one-parameter (strongly continuous) group of unitary operators $U(t) = e^{-iHt}$, where H is the Hamiltonian of the system. If a pure state $|\psi(0)\rangle$ describes a physical object which exists at time $t = 0$, then a history $|\psi(t)\rangle$ describes a persisting object which exists at each t and changes with time. The infamous measurement problem arises when we consider the relation of the unitary dynamics of the state to the

results of laboratory observations.

The *observables* of the system are self-adjoint operators[4] on $\mathcal{H}$ associated with measurable quantities, and the values they may take on measurement correspond to the *spectrum* of the operator. For an observable $\hat{A}$ with a discrete spectrum (*e.g.* the Hamiltonian of a simple harmonic oscillator), each spectral value a_n has an associated *eigenvalue* equation $\hat{A}|\phi_n\rangle = a_n|\phi_n\rangle$, where $|\phi_n\rangle$ is an *eigenstate* of $\hat{A}$. Distinct eigenvalues are associated with mutually orthogonal subspaces (*eigenspaces*) $|a_n\rangle$ which are spanned by the vectors $|\phi_n\rangle$ with zero inner product, $\langle\phi_m|\phi_n\rangle = 0$ for $m \neq n$. Any vector can be written as a weighted sum $|\psi\rangle = \sum_n |a_n\rangle\langle a_n|\psi\rangle = \sum_n c_n|\psi_n\rangle$, where $|\psi_n\rangle$ is the *projection* of $|\psi\rangle$ onto $|a_n\rangle$ and c_n are complex coefficients $\sum_n |c_n|^2 = 1$. This is known as the *spectral decomposition* or *resolution of the identity* of $\mathcal{H}$ with respect to $\hat{A}$, which we can write as $\hat{A} = \sum_n a_n|a_n\rangle\langle a_n|$.

According to the standard story, the probability of obtaining a particular value a_n in measurement is given by $\langle\psi|\psi_n\rangle = |c_n|^2$ (the *Born Rule*) and, having observed a system to take a particular value, upon repeating the measurement of the observable it will be found to have the same value a_n. However, according to the formalism the only way this could happen is if the system were in an eigenstate of $\hat{A}$ (known as the *eigenstate-eigenvalue link*), but since (i) in general a system is not in an eigenstate with respect to $\hat{A}$, and (ii) the dynamics provided by the Schrödinger equation are unitary, there is (in general) no reason to think that a system should ever be found in such a state. This is the measurement problem.

The third posit I will require is, therefore, that we consider *isolated* quantum systems which need only unitary evolution for their complete description over time; persisting quantum objects are isolated systems on this view. This means that we will not need to

[4]An operator $\hat{A}$ is symmetric on $\mathcal{H}$ *iff* $\langle\psi|\hat{A}\phi\rangle = \langle\hat{A}\psi|\phi\rangle$ for all elements in its domain $\mathcal{D}(\hat{A}) \subseteq \mathcal{H}$. It is self-adjoint $\hat{A} = \hat{A}^\dagger$ *iff* it is symmetric and $\mathcal{D}(\hat{A}) = \mathcal{D}(\hat{A}^\dagger)$.

concern ourselves with the measurement problem. This invites the worry that very few systems in the actual world will fall under this criterion. Maybe so, but on at least one interpretation of quantum mechanics (Everett-style realist 'no-collapse') *all* systems undergo only unitary evolution.

4. Parts and Spatial Parts What is a part of a quantum object? I contend that a suitable part-whole relation is provided by considering the subspaces of $\mathcal{H}$, or equivalently the projections onto those subspaces. According to classical mereology, the relation of *parthood* is (minimally) reflexive (everything is part of itself), transitive (if p is part of q and q is part of r then p is part of r) and antisymmetric (no two distinct things can be part of each other). As is well known, the subspaces of a vector space $A, B, C...$ are partially ordered by the relation of *inclusion*, which is reflexive ($A \subseteq A$), transitive (if $A \subseteq B$ and $B \subseteq C$ then $A \subseteq C$) and antisymmetric (if $A \subseteq B$ and $B \subseteq A$ then $A = B$).

I claim that in quantum mechanics the *spatial* parts are given in terms of the subspaces of the $\mathcal{H}$ associated with the spectral decomposition of the position observable $\hat{Q}$. As $\hat{Q}$ has a (purely) continuous spectrum, this will require some more details about self-adjoint operators. Since the pioneering work of von Neumann we have known that any self-adjoint operator (even an unbounded continuous operator) on $\mathcal{H}$ is uniquely associated (up to unitarity) with a *spectral measure* which allows us to replace the sum over projections onto eigenspaces with an integral $\hat{Q} = \int_{\mathbb{R}} \lambda \, dE_\lambda$, where E_λ is a spectral family of projections with $\lambda \in \mathbb{R}$. It is this which allows us to write the position operator as an integral over space $\hat{Q} = \int dx \, x \, |x\rangle\langle x|$.[5]

[5]However, note that the equation $\hat{Q}|q\rangle = q|q\rangle$ is a merely formal expression in this case *i.e.* $|q\rangle$ is a (so-called) *improper* eigenstate and not an element of $\mathcal{H}$.

We may associate with each Borel set[6] $\Delta \in \mathfrak{B}(\mathbb{R})$ a projection $P^{\hat{Q}}(\Delta) = \int_\Delta \lambda dE_\lambda$. The map $P^{\hat{Q}} : \Delta \mapsto P^{\hat{Q}}(\Delta)$ is known as a Projection Valued Measure (PVM) and has the properties (i) $P^{\hat{Q}}(\mathbb{R}) = 1$ (normalization), and (ii) $P^{\hat{Q}}(\bigcup_n \Delta_n) = \sum_n P_n^{\hat{Q}}(\Delta_n)$ (strong σ-additivity), where Δ_n is a sequence of mutually disjoint Borel sets $\Delta_m \cap \Delta_n = \emptyset$ for $m \neq n$. We can do this quite generally since the self-adjoint operators on $\mathcal{H}$ are in one-to-one correspondence with the set of PVM's (Teschl 2009, Thm. 3.7).

What does this have to do with spatial parts? Well, the Borel sets correspond to spatial regions in a very intuitive way since any two sets of spatial points which occupy the same volume of space are assigned the same Borel set (of $\mathbb{R}^3$ now); a Borel set is an *equivalence class* of sets of points under the relation *having the same volume*. Take an ordinary object that occupies exactly a cube. Pick four opposite vertexes of the cube which lie in a plane (such that not all four are on the same face). How many ways are there of dividing the cube into two parts of equal volume along that plane? Presumably we would want to say: "There is only one way, straight down the middle!" And this is just the answer we find from looking at the Borel sets.

However, if we consider instead the set of points that lie in the interior of the cube there are *three* ways: one that excludes the points that lie on the plane, one that gives them to the left hand part, and one that gives them to the right hand part.[7] The upshot of these sort of

[6] $\mathfrak{B}(\mathbb{R})$ is the smallest σ-algebra over $\mathbb{R}$ containing all open intervals of $\mathbb{R}$. A σ-algebra over a set is a (nonempty) collection of subsets closed under complementation and countable union.

[7] Also note that to allow spatial parts corresponding to non-measurable sets would open the door to paradoxical results regarding their re-composition, illustrated by the Banach-Tarski theorem. For a discussion of these issues see Arntzenius (2008).

considerations, I take it, is that we would rather associate spatial parts with Borel sets rather than sets of points (or, equivalently, only with sets of points dense in some open interval of $\mathbb{R}$).

So, if you accept the notion of parthood I articulated above, then the PVM $P^{\hat{Q}}$ associated with the position observable $\hat{Q}$ provides a neat assignment of spatial regions to parts of the state space. It has the attractive property that any spatial region is associated with a unique projection, and, furthermore, if two regions are disjoint $\Delta_1 \cap \Delta_2 = \emptyset$ then they are associated with mutually orthogonal projections $P^{\hat{Q}}(\Delta_1)P^{\hat{Q}}(\Delta_2) = P^{\hat{Q}}(\Delta_2)P^{\hat{Q}}(\Delta_1) = 0$, while if they overlap $\Delta_1 \cap \Delta_2 \neq \emptyset$ then their intersection has the unique projection $P^{\hat{Q}}(\Delta_1)P^{\hat{Q}}(\Delta_2) = P^{\hat{Q}}(\Delta_2)P^{\hat{Q}}(\Delta_1) = P^{\hat{Q}}(\Delta_1 \cap \Delta_2)$ (from property (ii)).

This is a *localization scheme* in the sense that performing a measurement that corresponds to a projection $P^{\hat{Q}}(\Delta)$ has the possible outcomes $\{0, 1\}$: either the system is located in Δ or the system is not located in Δ. Furthermore, these possibilities are mutually exclusive in that $P^{\hat{Q}}(\mathbb{R}/\Delta) = I - P^{\hat{Q}}(\Delta)$ (from (ii)). Therefore, the system may be said to be 'wholly located in Δ at t' on the condition that $P^{\hat{Q}}(\Delta)|\psi(t)\rangle = |\psi(t)\rangle$. Since in general the system will not be in an eigenstate of *any* projection $P^{\hat{Q}}(\Delta)$ we say not that it is located somewhere but rather that it is *localizable*. If a pure state $|\psi\rangle$ describes a quantum object then, I claim, the projections $P^{\hat{Q}}(\Delta)|\psi\rangle$ denote the spatial parts of the object.

Another characteristic of the PVM $P^{\hat{Q}}$ which justifies the contention that it provides an assignment of parts is that it *covaries* with spatial translations $U(a)^{-1}P^{\hat{Q}}(\Delta)U(a) = P^{\hat{Q}}(\Delta + a)$, where $U(a) = e^{-i\hat{P}a}$ is the one-parameter unitary group of spatial translations generated by the total momentum $\hat{P}$. Roughly, this is a consequence of the fact that $\hat{Q}$ and $\hat{P}$ are canonically conjugate $[\hat{Q}, \hat{P}] = i\hbar$. Viewing these transformations passively as a relabeling of the spatial axis, *covariance* assures us that we are picking out the same parts despite having changed their relation to the labels.

Now, there is something a little disconcerting about these spatial parts. Firstly, (in general) each quantum object appears to be composed of parts that together cover all of space (from (i)). Secondly, these spatial parts do not 'move with the object' since $U(t)^{-1}P^{\hat{Q}}(\Delta)U(t) = P^{\hat{Q}}(\Delta)$. Neither of these features represent genuine problems for this view. First, there is nothing metaphysically necessary about the view that physical objects have limited spatial extent. Fields, for example, qualify as genuine physical entities without being limited to a particular region of space. Moreover, although the localization scheme necessarily covers all of space, the object itself may be localized in the above sense. Second, this might be thought of as a boon for the perdurantist since it restores a symmetry between time and space by removing the need to define spatial parts relative to spatial *location* (see Butterfield (1985)).

Another potentially disconcerting feature of these quantum spatial parts is that they are defined in terms of the position observable for the *entire* system with state $|\psi\rangle$, and so it may be the case that even though we (naively) suppose the system to be further decomposed into distinct subsystems $|\psi\rangle = |\eta\rangle \otimes |\xi\rangle$, the spatial parts assigned in this way fail to respect this decomposition such that the spatial degrees of freedom of the subsystems fail to be independent. This is known as *entanglement*, and is a pervasive feature of quantum mechanics. If the subsystems are considered to be spatially separated then entanglement may result in non-locality, in the sense that the results of local measurements of position on one subsystem may depend on the results of local (but distant) measurements on the other subsystem. The view taken here is that this apparent tension results from an incorrect notion of mereology: subsystems do not correspond to independent spatial parts *unless* they are associated with mutually orthogonal projections $P^{\hat{Q}}(\Delta)$.

5. (No) Temporal Parts What is a *temporal* part of a quantum object? A possible response might go as follows: Since the instantaneous quantum state determines all the kinematical properties of the system, we can specify the temporal parts of a quantum object by a simple assignment of the states $|\psi(t)\rangle$ to times $t \in \mathbb{R}$. This 'naive' scheme would assign to arbitrary sets of times $\{T\}$ the temporal part $|\psi(t)\rangle$ only if $t \in \{T\}$. However, the scheme completely fails to provide a partition of the object into *parts*. The problem is that the parameter t indexes a family of temporal *translations*, so fails to respect the requirement that temporal parts be 'wholly present' at t.

By means of analogy, consider the family of states $|\psi(a)\rangle = U(a)|\psi(0)\rangle$ where $U(a)$ is again the group of spatial translations by a. The naive spatial location scheme would assign parts (subspaces of $\mathcal{H}$) to spatial points $\{X\}$ according to whether or not value of the index a lies in $\{X\}$. But the claim that $|\psi(a)\rangle$ is 'wholly located' at the position a doesn't make sense since a merely denotes the spatial interval by which the state $|\psi(0)\rangle$ was translated. In general $|\psi(0)\rangle$ will not be located anywhere in particular (unless an eigenstate of some $P^{\hat{Q}}(\Delta)$) and so $|\psi(a)\rangle$ picks out the same 'part' as $|\psi(0)\rangle$.

Exactly the same analysis applies to the temporal translations $U(t)$. So to identify distinct temporal *parts* the perdurantist needs a temporal localization scheme which assigns to temporal intervals (proper) subspaces of the state space of the system, and so parts of $|\psi\rangle$: that is, a Projection Valued Measure $P^{\hat{T}}(\Delta)$ associated with a self-adjoint operator $\hat{T}$. In order that this scheme picks out genuine temporal parts, we should expect this scheme to *covary* with time translations $U(t)^{-1}P^{\hat{T}}(\Delta)U(t) = P^{\hat{T}}(\Delta + t)$ so that under a relabeling of the time axis the labels change but not the parts.

Unfortunately for the would-be quantum perdurantist, it turns out that these requirements are in conflict with the restriction on physical Hamiltonians known as the *spectral condition*,

which permits only Hamiltonian operators with a spectrum bounded from below *i.e.* only systems with a state of lowest energy. The usual argument for this is that to do otherwise would allow for systems which may transfer energy to their surroundings indefinitely. While it is true that all systems we know obey the spectral condition (*e.g.* a free particle or harmonic oscillator), we could also view it as a principle of the theory on par with the second law of thermodynamics.

Now, it is a theorem that if a self-adjoint Hamiltonian on $\mathcal{H}$ obeys the spectral condition then there can be no time PVM that covaries with time translations (see Srinivas and Vijayalakshmi (1981, Thm. 1) or Halvorson (2010)). Roughly, the spectral condition implies that any two vectors in $\mathcal{H}$ related by a time translation are non-orthogonal, so that the only assignment of temporal intervals to mutually orthogonal subspaces is $P^{\hat{T}}(\Delta) = 0$ for all Δ.[8] Thus no quantum object has (proper) temporal parts.

It is worth emphasizing that the problem is not that we cannot find a covariant assignment of temporal intervals to operators, but rather that there is no such assignment to projections on $\mathcal{H}$. So while it *is* the case that we can find a covariant mapping of intervals to operators in the form a Positive Operator Valued Measure (POVM), these assignments come without an associated spectral decomposition of $\mathcal{H}$ and, moreover, are non-unique (Hegerfeldt and Muga 2010). This failure to find a temporal decomposition of the state space into distinct subspaces means that a quantum object cannot have temporal parts in the same way as it has spatial parts.

However, such POVM's do provide a *generalized* resolution of the identity parameterized by t so giving meaning to the notion of a temporal interval in terms of operators on $\mathcal{H}$ (Holevo 1982). Therefore I see no reason to deny that time should be afforded the status of a physical parameter – although not one associated uniquely with a self-adjoint operator – and in this

[8]This implies there is no self-adjoint operator $\hat{T}$ canonically conjugate to the Hamiltonian.

sense I disagree with Halvorson's claim that "time [in quantum theory] is not a quantity at all – not even an unobservable quantity" (Halvorson 2010, 1).

6. Conclusion For the perdurantist, the failure to find a temporal localization scheme has a worrying implication for the claim that persisting quantum objects have temporal parts: If an object has temporal parts then times should be associated uniquely with subspaces of $\mathcal{H}$ just as spatial regions are uniquely associated with spatial parts through a localization scheme $P^{\hat{Q}}(\Delta)$. The claim that the part-whole relation applied to space is the same as the part-whole relation applied to time is demonstrably false when applied to quantum objects. To the extent that we have reason to think that all persisting objects are quantum objects, this provides reason to doubt that perdurantism is true.[9]

In fact, the result we have demonstrates that only two temporal partitions are consistent with the requirements: Either there are no temporal parts, or there is one part corresponding to the entire history $|\psi(t)\rangle$. While I have argued that both these options are problematic for the perdurantist, the former is consistent with endurantism since the endurantist maintains that persisting objects have no temporal parts and no temporal width; the endurantist claims there is exactly one persisting object existing at each moment, and so may consistently attribute to that object at time t the state $|\psi(t)\rangle$.

[9]Arguably, classical persisting objects are best thought of as "patterns that emerge from an ubiquitous, continuous, and very efficient process of decoherence." Butterfield (2006, 41). Decoherence refers to the process by which interactions between an 'object' system (*e.g.* a dust particle) and its environment serve to pick out a dynamically 'preferred' basis according to which the object system is approximately diagonalized. My argument concerns the basis independent description of the entire system of object *and* environment.

However, the latter option admits a valid four-dimensional interpretation which I call *temporal holism*, corresponding to the idea that the quantum state has exactly one temporal part comprising its entire history.[10] This offers an interesting resolution of the problem of persistence since it effectively denies that persisting objects change with time. It is distinct from endurantism in that although the same object is present at each time it is never wholly present; and distinct from perdurantism in the sense that although the persisting object exists at many times, no part of it is ever wholly present either.

A similar view has been advocated by Rovelli (2004) on the basis of relativistic considerations that he traces back to Dirac's preference for the Heisenberg formulation of quantum mechanics (in which the observables not the state are regarded as varying in time) over the Schrödinger picture (adopted above, in which the state varies not the observables). Since the Schrödinger and Heisenberg pictures are strictly equivalent within non-relativistic quantum mechanics the situation there is effectively neutral with respect to temporal holism and endurantism. Nonetheless, temporal holism may be thought to win out to the extent that four-dimensionalism is encouraged by relativity, having ruled out perdurantism by the above argument.

[10]This is similar to the 'worm view' advocated by Balashov (2010) in the context of special relativity but there are obvious difficulties with describing quantum systems in terms of world-tubes.

References

Arntzenius, F. (2008). Gunk, topology and measure. In D. Zimmerman (Ed.), *Vol. 4*, Oxford Studies in Metaphysics, pp. 225–247. Oxford: Oxford University Press.

Balashov, Y. (2010). *Persistence and spacetime*. Oxford: Oxford University Press.

Butterfield, J. (1985). Spatial and temporal parts. *The Philosophical Quarterly 35*(138), 32–44.

Butterfield, J. (2005). On the persistence of particles. *Foundations of Physics 35*(2), 233–269.

Butterfield, J. (2006). The rotating discs argument defeated. *The British Journal for the Philosophy of Science 57*(1), 1–45.

Galapon, E. (2002). Pauli's theorem and quantum canonical pairs. *Proceedings of the Royal Society of London A 458*(2018), 451–472.

Halvorson, H. (2010). Does quantum theory kill time? http://www.princeton.edu/hhalvors/papers/notime.pdf.

Hawley, K. (2004). *How things persist*. Oxford: Clarendon Press.

Hegerfeldt, G. C. and J. G. Muga (2010). Symmetries and time operators. *Journal of Physics A 43*(50), 1–18.

Holevo, A. (1982). *Probabilistic and Statistical Aspects of Quantum Theory*. Amsterdam: North-Holland.

Lewis, D. K. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

McCall, S. and E. J. Lowe (2003). 3D/4D equivalence, the twins paradox and absolute time. *Analysis 63*, 114–123.

Nairz, O., M. Arndt, and A. Zeilinger (2003). Quantum interference experiments with large molecules. *American Journal of Physics 71*, 319.

Rovelli, C. (2004). *Quantum Gravity*. Cambridge: Cambridge University Press.

Sider, T. (1997). Four-dimensionalism. *The Philosophical Review 106*(2), 197–231.

Srinivas, M. and R. Vijayalakshmi (1981). The 'time of occurrence' in quantum mechanics. *Pramana 16*(3), 173–199.

Teschl, G. (2009). *Mathematical methods in quantum mechanics*. Providence: American Mathematical Society.

Unruh, W. G. and R. M. Wald (1989). Time and the interpretation of canonical quantum gravity. *Physical Review D 40*(8), 2598–2614.

It's Okay to Call Genetic Drift a "Force"

Charles H. Pence

Preprint, July 25, 2012

Abstract

One hotly debated philosophical question in the analysis of evolutionary theory concerns whether or not evolution and the various factors which constitute it (selection, drift, mutation, and so on) may profitably be considered to be "forces" in the traditional, Newtonian sense. Several compelling arguments assert that the force picture is incoherent, due to the peculiar nature of genetic drift. I consider two of those arguments here – that drift lacks a predictable direction, and that drift is constitutive of evolutionary systems – and show that they both fail to demonstrate that a view of genetic drift as a force is untenable.

1. Introduction

The evolution of populations in nature is described in many ways, using a whole host of smaller factors with extensive theories of their own: natural selection, genetic drift, mutation, migration, linkage disequilibrium, meiotic drive, extinction, increase in complexity, and so on. The natural philosophical question, then, is this: what is the relationship between these "component" theories and the overall trajectory of evolution in the broad sense?

Work on this question has recently focused on the *causal* picture implied by this relationship. Is evolution (as a whole) a causal process? Do some of the smaller-scale theories describe causal processes? Which ones? And how do those smaller-scale causal processes combine to produce the resultant trajectory of populations through time? Two positions on these questions have crystallized. One, the "statisticalist" interpretation of evolutionary theory (e.g., Walsh et al., 2002; Matthen and Ariew, 2002), claims that both evolution as a whole and these smaller-scale theories do not describe causal processes. Rather, the causal processes at work exist at the level of individual organisms and their biochemistry: individual instances of survivals, deaths, predations, mutations, and so forth. All these theories, then, constitute quite useful, but *not* causal, ways in which we may statistically combine events to enable us to grasp interesting trends within populations of causally interacting individuals.

The other view, the "causalist" interpretation (e.g., Millstein, 2002, 2006; Shapiro and Sober, 2007), considers all of these processes to be genuinely causal.

Evolution causes changes in populations, as do selection, mutation, migration, genetic drift, and so forth. How exactly we specify these causal processes varies – for example, as different varieties of "sampling" (Hodge, 1987), as population-level causes (Millstein, 2006), or as supervening on lower-level causes (Shapiro and Sober, 2007) – but they are causal nonetheless.

This heated debate has produced much work on an allied problem which will be the topic of my discussion here. It is a common pedagogical trope in the teaching of biology to describe all of these smaller-scale theories as referring to *forces,* each of which propels a population in a different direction through some space (of morphologies, phenotypes, genotypes, etc.) with a different strength, adding together in some sense to produce the population's overall evolutionary trajectory over time. Crow and Kimura introduce a discussion of equilibrium under selection pressure by noting that "ordinarily one regards selection as the strongest force influencing gene frequencies" (1970, p. 262). Hartl and Clark discuss the possibility of balancing mutation and drift, writing that "there are many forces in population genetics that act in opposition to one another, and it is this tension that makes for interesting behavior at the population level. […] Merely because these two forces are in opposition, it does not guarantee that there will be a stable balance between them" (1997, p. 294). Strickberger argues that since mutational equilibrium is not reached in many natural populations, "other forces must be responsible for the establishment of gene frequencies" (1968, p. 719). This pedagogical pattern is even common at the high school level: in a chapter titled "The Forces of Evolutionary Change," Lewis summarizes natural selection, nonrandom mating, mutation, migration, and genetic drift in a force-like diagram (1997, p. 412).

I have quoted from several textbooks to demonstrate the pervasiveness of this 'force' metaphor at all levels of biological pedagogy. But what of it? Why is this particular biological turn of phrase of philosophical interest? In his original introduction of what would become the causalist interpretation, Sober (1984) described, influentially, evolutionary theory as a *theory of forces.* Sober's metaphor is intended to carry some genuine explanatory weight. Allowing, of course, that the analogy here is not entirely precise, he claims that *just as* component, causal forces are summed together to determine the net force acting on a body in Newtonian dynamics, a force-like understanding is the right way to picture not just the metaphorical structure of evolution, but its *causal* structure as well. Sober writes that in addition to work on the history of life,

> evolutionary biology has also developed a theory of *forces.* This describes the *possible causes* of evolution. The various models provided by the theory of forces describe how a population will evolve if it begins in a certain initial state and is subject to certain causal influences along the way. (Sober, 1984, p. 27)

This view makes evolution, in the apt terminology deployed by Maudlin, a "quasi-Newtonian" theory (2004, p. 431). "There are, on the one hand, *inertial* laws that describe how some entities behave when nothing acts on them, and then

there are laws of *deviation* that specify in what conditions, and in what ways, the behavior will deviate from the inertial behavior" (Maudlin, 2004, p. 431). This is, Maudlin notes, a very natural way for us to understand the behavior of systems: whether or not the laws of a given system are amenable to such analysis, we *like* to produce quasi-Newtonian theories.

But to deploy force language in this more substantive way brings Sober in for another line of argument in addition to the critiques aimed at the causal view in general.[1] For we now must ask about the soundness of this appropriation of Newtonian force. Should selection and drift be treated in this way, or not? One recurring difficulty with adopting the force metaphor is the issue of genetic drift. A common refrain in this debate claims that considering drift to be similar to a Newtonian force is highly problematic.

In what follows, I will argue in favor of the force metaphor, by taking on two arguments against the tenability of considering drift as a force. The first is the (by now, well-trodden) claim that genetic drift, though its magnitude may be determined by the effective population size, lacks a direction specifiable or predictable in advance. Since all Newtonian forces, it is said, must have specifiable magnitudes *and* directions, drift cannot be considered a force, and the metaphor thus falls apart. The second argument claims that it is a category mistake to consider drift a force which impinges upon populations. It is, rather, the default state in which populations find themselves. All evolving populations *necessarily* drift, and thus to describe drift as an "external" force is misleading. Both of these critiques, I will show, miss the mark.

2. The Direction of Drift

It is by now an old chestnut in this debate that genetic drift lacks a specifiable or predictable direction. Matthen and Ariew (2002, p. 61) note in a dismissive aside that "in any case, drift is not the sort of thing that can play the role of a force – it does not have predictable and constant direction." Brandon (2006) adopts the same argument, and it is one of the central motivations behind his development of the "zero-force evolutionary law" (Brandon, 2006, 2010; McShea and Brandon, 2010).

The basic outline is straightforward. Genetic drift, often called "random" drift, is a stochastic process. Consider a population which is uniformly heterozygous for some allele Aa – all members of the population possess one copy of the dominant allele (A) and one copy of the recessive allele (a). Assuming no selection, mutation, or other evolutionary forces act on the population, genetic drift will eventually drive this population toward homozygosity, uniformity at either AA or aa, with one of the two alleles removed from the population. This

1. Early in the debate between causalists and statisticalists, this point was often missed – Matthen and Ariew (2002), for example, take it to be a point against *the causal interpretation itself* that genetic drift cannot be described as a force. This entails, at best, that the force metaphor should be discarded, not that the causal interpretation is untenable, a point stressed by Stephens (2004) and Millstein (2006).

is because the homozygous states AA and aa are what we might call "absorbing barriers" – once a population has lost all of its A or a alleles (and again, given that there is no mutation), it is "stuck" at the uniform homozygous state. The "random walk" of genetic drift will, given enough time, eventually arrive and remain at one or the other of these permanent states.

Here, then, is the rub – the population will arrive at *one* of these states, but it is impossible in advance to predict which one will be its eventual fate. In this sense, at least, the population-level outcome of genetic drift is random.[2] It is obvious, the argument concludes, that drift cannot act as a Newtonian force, because Newtonian forces have directions that may be specified and predicted. Consider natural selection. The direction in which selection will drive a population is obvious, and is indeed specifiable in advance: selection will move populations in the direction of increased fitness. We may even visualize the "adaptive landscape" in the absence of any actual populations, specifying the direction of the selective force prior to any actual population's experiencing it.[3] Such analysis is clearly impossible for drift, and drift cannot therefore be described as a force.

Two responses on behalf of the force metaphor have been offered. In our initial discussion of drift above, drift was described fairly clearly in directional terms: it drives populations toward homozygosity (Stephens, 2004, pp. 563–564). Insofar as this is a *direction,* we may avoid the objection. There are several reasons that we might be worried about this response, however. First, Filler has argued persuasively that if we are *too* liberal with our force metaphor, we run the risk of sapping the notion of 'force' of all its explanatory power. Consider, for example, Molière's classic satire of opium's "dormitive virtue." We could construct a "fatigue-space" in which sleep sits at the end of one axis, and then describe a "dormitive force" which drives persons up the sleep axis. Ascribe this "dormitive force" to opium, and we have come close to completing Molière's folly, providing a nearly empty "explanation" for opium's causing sleep (Filler, 2009, pp. 779–780). If "heterozygosity-space" resembles "fatigue-space" in Filler's sense too closely, then the "toward homozygosity" response to this objection fails.

Another worry about "toward homozygosity" as a direction for drift is that it may mischaracterize what it is that drift is intended to describe. As mentioned above, drift has a direction toward homozygosity insofar as (in the absence of mutation and migration) homozygosity constitutes a set of absorbing barriers for the state of a population. What drift is genuinely about, however, is not the existence of these barriers – which are set by the mutation and migration constraints – but rather the population's behavior *between* these barriers. This "toward homozygosity" direction of genetic drift, therefore, is not a feature of drift itself, but defined by other parts of evolutionary theory; thinking that

2. The sense of "stochastic" and "random" at work here is, therefore, a subjective one. Whether or not there exists a stronger type of stochasticity underlying genetic drift, and what exactly this sense might amount to, seems to hinge in large part on the result of the debate over drift's causal potency (see Rosenberg, 2001).

3. Though see Pigliucci and Kaplan (2006) for some of the difficulties with the adaptive landscape metaphor.

"toward homozygosity" is a feature *of* drift thus may be mistaken.

We have several independent reasons, then, for suspecting that the defense of the force view by appeal to drift's direction "toward homozygosity" is problematic. If this is true, we must look for another way to resolve the trouble with drift's direction, and the second available response turns to the definition of 'force' itself. Perhaps the trouble with the objection lies in its rigorous adherence to the claim that forces must have directions predictable in advance.[4] Could we discard this requirement *without* discarding the extra explanatory power that the notion of a 'force' provides us?

One attempt to do so is offered by Filler (2009, pp. 780–782). He argues that we may harvest two specific criteria for forces from the literature on Newtonian systems: namely, that forces be both *precisely* numerically specifiable in magnitude and able to unify our explanations of a large array of phenomena. Such criteria, it is presumed (though not argued), would forestall the "dormitive force" while permitting genetic drift. Even if they do not, however, Filler notes that "we could still posit a continuum of forces with maximally precise and unifying forces on one end and mathematically vague and weakly unifying forces on the other" (Filler, 2009, p. 781).

What of this attempt to salvage the force view? In general, I am broadly sympathetic with the response of carefully weakening the criteria for 'force'-hood. I would like, however, to support the same conclusion by a slightly different line of argument. While the literature that Filler cites to establish mathematical specifiability and unifying power as desiderata for forces is valuable, I am concerned about it for two reasons. First, given that these criteria are offered by Filler without providing an analysis of genetic drift or any other forces, they seem dangerously close to being ad-hoc additions to our force concept. Is there a principled argument for why these criteria should replace that of directionality, in general? Second, Filler does not offer a direct argument that genetic drift passes these criteria, so we can't yet be sure that the argument he provides gives us the result that we're looking for. I believe both of these deficits can be remedied by comparing genetic drift to a different force that is standardly invoked in Newtonian dynamics: Brownian motion.

2.1. Brownian Motion

My claim, then, is this: whatever our general analysis of a force winds up being, it happens to be the case that we *already* countenance examples of forces that do, indeed, have stochastically specified directions, namely, the force of Brownian motion. This argument is admittedly less ambitious than that of Filler – we do not, for example, wind up with enough theoretical resources to fully specify the continuum from paradigm cases of forces to fringe cases. But we do have precisely what we need to countenance genetic drift as a force, for genetic drift,

4. The claim that forces must have specifiable directions appears, at least, in Matthen and Ariew (2002); Stephens (2004); Brandon (2005); and Brandon (2006).

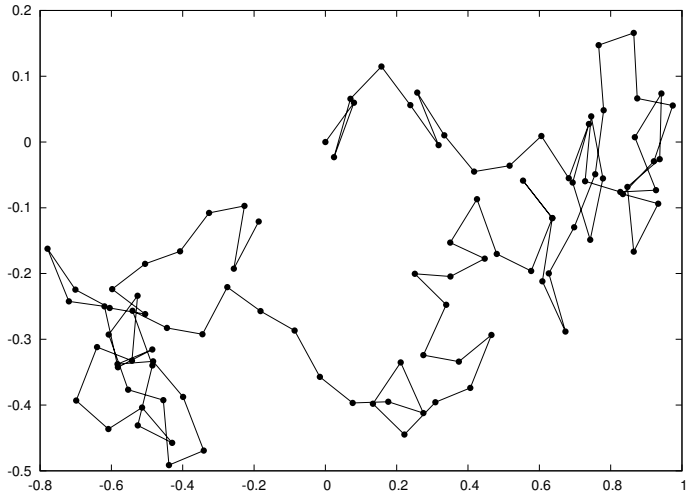


Figure 1: A simulation of a particle released at $(0, 0)$ undergoing Brownian movement. Inspired by Perrin's drawing of the Brownian motion of colloidal particles in water, viewed under the microscope (fig. 6 of Perrin, 1909, p. 81).

it turns out, can be formulated precisely analogously to the force of Brownian motion.

Brownian motion is a common occurrence. The behavior of dust particles as they float through a sunny window or a glass of water is governed in large part by the manner in which they collide with the molecules of the fluid in which they are suspended (see Figure 1). Since the motion of the fluid molecules is itself modeled stochastically (with the tools of statistical mechanics), it is unsurprising that Brownian motion in turn is a stochastic force.

What does the formal representation of a stochastic classical force look like? The now-standard derivation of the mathematics of Brownian motion was provided by Langevin in 1908 (translated in Lemons and Gythiel, 1997):

$$m\frac{d^2x}{dt^2} = -6\pi\mu a\frac{dx}{dt} + X. \tag{1}$$

This is a stochastic differential equation, with x representing the location of the particle within the fluid, m its mass, a damping coefficient $-6\pi\mu a$ (which describes the manner in which the viscosity of the fluid through which the particle moves slows its travel), and a random "noise term" X, which describes the actual effect of the collisions with fluid molecules.

A few observations about this equation are in order. First, it is written as an equation for a force: $m \cdot d^2x/dt^2$ is just mass times acceleration, so we could equivalently have written $F = -6\pi\mu a \cdot dx/dt + X$. Nor need one quibble that the differential equation specifying this force references the particle's velocity, dx/dt. Equations for many other forces do so as well, including friction in air or

water (drag). Secondly, the "source" of the randomness here is obvious, coming entirely from the noise term X. About it, Langevin says that "we know that it is indifferently positive and negative and that its magnitude is such that it maintains the agitation of the particle, which the viscous resistance would stop without it" (Lemons and Gythiel, 1997, p. 1081).

Finally, the force described by this equation bears all of the same "problematic" characteristics as genetic drift. Most importantly, its direction can by no means be predicted in advance: nothing about the direction of the force described by equation (1) is "determinate" in this sense. It depends entirely on the noise term which, as Langevin notes, "indifferently" (that is to say, randomly) changes sign and magnitude as the system evolves. The same is, of course, true of genetic drift, under which an allele frequency is equally likely to increase or decrease at each point in time. The example of Brownian motion, therefore, offers us a case in which the notion of 'force' is weakened in *precisely* the way required to countenance genetic drift – by admitting forces that vary in direction stochastically over time.

The opponents of the force view still have one obvious way to respond to this argument. They might reject outright the extension of force talk to both Brownian motion and genetic drift. While this is a perfectly coherent choice, I am not certain what the motivation for it would be. Of course, when we introduce a stochastic force, we introduce an element of unpredictability into our system, rendering null one of the primary benefits of a classical, force-based picture: the ability to use information about component force values to make determinate advance predictions about the behavior of systems. But we already lack the ability to make such detailed predictions of individual biological systems – why would we think that a force-based view of evolutionary theory would somehow make them possible? The question, rather, is simply whether it is possible to maintain a "net-force" picture of evolutionary theory which includes the randomness of genetic drift, and the example of Brownian motion shows this to be clearly achievable, should we be inclined to do so.

Further, just because the values are not predictable in advance does not mean that these stochastic forces somehow cannot be taken into account in the development of models. The Wright-Fisher model of genetic drift has spawned much research in population genetics as a computational/mathematical model of the action of genetic drift, and, similarly, Brownian motion can be taken into account in models of fluid dynamics when it is taken to be an important factor (see, e.g., Huilgol and Phan-Thien, 1997).

Finally, it seems that many authors in the debates over the causal structure of evolution either explicitly tolerate or make room for forces of different sorts such as these. McShea and Brandon, for example, when discussing how we might arrive at the "correct" distribution of evolutionary causes into forces, note their skepticism that "there are objective matters of fact that settle what counts as forces in a particular science, and so what counts as the zero-force condition" (2010, p. 102). That is, while facts can settle what causal influences are at work in a given system, they cannot, according to McShea and Brandon, settle how we

partition these causal processes into "forces." Even the statisticalist analysis of Walsh, Lewens, and Ariew describes as a paradigm case of Newtonian, dynamical explanation the case of a feather, "affected not only by the force of gravity but also by attractive forces from other bodies, electromagnetic forces, *forces imparted by random movements of the air molecules,* etc." (2002, p. 454, emph. added). I claim that without further argument, there is little reason to dogmatically adhere to the requirement that forces have directions specifiable in advance.

3. Drift as "Constitutive" of Evolutionary Systems

Another line of attack on the force view, marshaled by Brandon, doesn't turn on the appropriateness of stochastic-direction forces. Rather, it claims that it is a category mistake (or something close to it) to consider drift as an *external* force that acts on biological systems. Drift, on the contrary, is "part and parcel of a constitutive process of any evolutionary system," and is therefore *necessarily* found in any set of circumstances in which evolution is possible. "Force" talk, on the other hand, should be reserved for forces which appear in "special" circumstances. In the biological case, mutation, selection, and migration (among others) are "special" forces, but drift, as a "constitutive" component of evolution, is not – it is part of the "zero-force" state of evolutionary systems (Brandon, 2006, p. 325).

To help elucidate this argument further, return to Maudlin's discussion of "quasi-Newtonian" systems as mentioned in the introduction (2004, p. 431). Maudlin points out a very valuable psychological or motivational distinction between our inertial or zero-force laws and our deviation or force laws. Namely, the zero-force conditions are supposed to be what influences a body when, in some particularly relevant sense, *nothing is happening to it.* The appropriate sense of "nothing happening" is obviously domain-relative, and Brandon's claim seems to be precisely that placing drift on the side of the force laws is a poor definition of "nothing happening." When nothing is happening to a biological system, he argues, *it drifts.*

Again, let's turn to an analogy with classical mechanics. Classical mechanics has its own set of highly pervasive forces, and for each of these we have made the implicit decision to consider that force not as part of the inertial conditions, but as a deviation from those conditions. Take gravitation, for example. We might reply to Brandon's objection that gravitation is as universal in Newtonian systems as genetic drift is in evolutionary systems. Applying the logic of Brandon's objection here, then, Newton's first law is incorrectly formulated. Gravitation should be considered part of the "default" or "zero-force" state of Newtonian mechanics. While this isn't an outright reductio, it strikes me that any discussion of forces which fails to handle the paradigm case of Newtonian gravitation is seriously flawed.

I suspect, however, that the supporter of this objection would reply that there is an important and salient difference between genetic drift and gravitation.

While there may be no Newtonian system which *in fact* exhibits no gravitational effects, it is possible to describe in Newtonian terms a system that would not be subject to gravitation – either by dialing the gravitational constant G back to zero, or by imagining the behavior of an isolated test mass "at infinity," infinitely distant from all other mass in the universe. Gravitation therefore is not *necessary* for the description of a Newtonian system in the way that drift is for an evolutionary system.

It is not obvious to me, however, that there is any conceptual difficulty in abstracting genetic drift away from an evolutionary system. Imagine an infinite population with individuals initially equally distributed among four possible genotypes, A, B, C, and D. Parents produce offspring identical to themselves, modulo a small mutation rate. There exists a selective force, which causes types C and D to have a 10% chance of dying before reaching reproductive age. Finally, the reproductive output of each type in the next generation is set in advance: say that all types produce exactly one offspring if they survive to reproductive age, and then die. Here we have an example of a thought experiment on which selection exerts an influence (types C and D will clearly eventually die out), mutation has an influence (due to the non-zero mutation rate), but genetic drift has none. The population is infinite, so we have no bottleneck effects or effects of finite population size. Further, each individual has a guaranteed reproductive outcome from birth, based upon its type – and to the extent that these outcomes are probabilistic, this is the influence of *selection* or *mutation,* not *drift.* Indeed, we can predict that in the infinite limit, the population will consist of roughly half A organisms and half B.[5]

Is there anything more outlandish about this drift-free toy model than an example consisting of a universe containing only one isolated and non-extended point mass, free of gravitation, or a test mass at infinite distance from all other masses? Clearly there are no infinite populations in the real world, but here it seems we have a perfectly tenable thought-experiment on which we may separate the effect of drift from all the other evolutionary forces, and then reduce that effect to zero. There is nothing any more "constitutive" about drift for evolutionary systems than there is about gravitation for Newtonian systems.

4. Conclusion

I have here considered two arguments against the conceptual tenability of considering genetic drift as a "force" like those of Newtonian dynamics. The first asserted that genetic drift lacks a predictable direction. This argument fails by virtue of an analogy with Brownian motion: if Brownian motion is a satisfactory force (and, I have argued, it is), then so is genetic drift. The second argument against drift-as-force proposed that drift is a constitutive feature of evolutionary systems. This argument fails because accepting its premises results in a misun-

5. With a small, but predictable, fraction of newly-arisen mutants. I am indebted to Grant Ramsey's thoughts on drift for helping me devise this example.

derstanding of the relationship between Newtonian gravitation and inertia.

I have, of course, done nothing here to resolve the overall debate between the causal and statistical interpretations of evolutionary theory. But the utility of the force metaphor in the description of evolutionary systems makes it something worth defending – and it continues to survive the host of objections raised against it.

Bibliography

Brandon, R.N. 2005. *The difference between selection and drift: A reply to Millstein*. **Biology and Philosophy**, 20(1):153–170. doi: 10.1007/s10539-004-1070-9.

———. 2006. *The principle of drift: biology's first law*. **Journal of Philosophy**, 103 (7):319–335.

———. 2010. *A non-Newtonian model of evolution: the ZFEL view*. **Philosophy of Science**, 77(5):702–715.

Crow, J.F. and M. Kimura. 1970. *An introduction to population genetics theory*. Blackburn Press, Caldwell, NJ.

Filler, J. 2009. *Newtonian forces and evolutionary biology: a problem and solution for extending the force interpretation*. **Philosophy of Science**, 76:774–783.

Hartl, D.L. and A.G. Clark. 1997. *Principles of population genetics*. Sinauer Associates, Sunderland, MA, 3rd edition.

Hodge, M.J.S. 1987. *Natural selection as a causal, empirical, and probabilistic theory*. In Krüger, L., G. Gigerenzer, and M.S. Morgan, editors, **The probabilistic revolution**, pages 233–270. The MIT Press, Cambridge, MA.

Huilgol, R.R. and N. Phan-Thien. 1997. *Fluid mechanics of viscoelasticity*. Elsevier Science B.V., Amsterdam.

Lemons, D.S. and A. Gythiel. 1997. *Paul Langevin's 1908 paper "On the theory of brownian motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530-533 (1908)]*. **American Journal of Physics**, 65(11):1079–1081.

Lewis, R. 1997. *Life*. WCB/McGraw-Hill, Boston, MA.

Matthen, M. and A. Ariew. 2002. *Two ways of thinking about fitness and natural selection*. **Journal of Philosophy**, 99(2):55–83.

Maudlin, T. 2004. *Causation, counterfactuals, and the third factor*. In Collins, J.D., N. Hall, and L.A. Paul, editors, **Causation and counterfactuals**, pages 419–443. The MIT Press, Cambridge, MA and London.

McShea, D.W. and R.N. Brandon. 2010. *Biology's first law: the tendency for diversity and complexity to increase in evolutionary systems*. University of Chicago Press, Chicago and London.

Millstein, R.L. 2002. *Are random drift and natural selection conceptually distinct?* **Biology and Philosophy**, 17:33–53.

———. 2006. *Natural selection as a population-level causal process*. **British Journal for the Philosophy of Science**, 57(4):627–653. doi: 10.1093/bjps/axl025.

Perrin, J.B. 1909. *Mouvement brownien et réalité moléculaire*. **Annales de chimie et de physique**, VIII(18):5–114.

Pigliucci, M. and J.M. Kaplan. 2006. *Making sense of evolution: the conceptual foundations of evolutionary theory*. University of Chicago Press, Chicago.

Rosenberg, A. 2001. *Discussion note: indeterminism, probability, and randomness in evolutionary theory*. **Philosophy of Science**, 68(4):536–544.

Shapiro, L. and E. Sober. 2007. *Epiphenomenalism – the do's and the don'ts*. In Wolters, G. and P. Machamer, editors, **Thinking about causes: from Greek philosophy to modern physics**, pages 235–264. University of Pittsburgh Press, Pittsburgh, PA.

Sober, E. 1984. *The nature of selection*. The MIT Press, Cambridge, MA.

Stephens, C. 2004. *Selection, drift, and the "forces" of evolution*. **Philosophy of Science**, 71(4):550–570. doi: 10.1086/423751.

Strickberger, M.W. 1968. *Genetics*. Macmillan and Co., New York.

Walsh, D.M., T. Lewens, and A. Ariew. 2002. *The trials of life: natural selection and random drift*. **Philosophy of Science**, 69(3):429–446. doi: 10.1086/342454.

Why internal validity is not prior to external validity

Johannes Persson & Annika Wallin
Lund University, Sweden
Corresponding author: johannes.persson@fil.lu.se

[**Abstract:** We show that the common claim that internal validity should be understood as prior to external validity has, at least, three epistemologically problematic aspects: experimental artefacts, the implications of causal relations, and how the mechanism is measured. Each aspect demonstrates how important external validity is for the internal validity of the experimental result.]

1) Internal and external validity: perceived tension and claimed priority

Donald T. Campbell introduced the concepts internal and external validity in the 1950s. Originally designed for research related to personality and personality change, the use of this conceptual pair was soon extended to educational and social research. Since then it has spread to many more disciplines.

Without a doubt the concepts captures two features of research scientists are aware of in their daily practice. Researchers aim to make correct inferences both about that which is actually studied (internal validity), for instance in an experiment, and about what the results 'generalize to' (external validity). Whether or not the language of internal and external validity is used in their disciplines, the tension between these two kinds of inference is often experienced.

In addition, it is often claimed that one of the two is prior to the other. And the sense in which internal validity is often claimed to be prior to external validity is both temporal and epistemic, at least. For instance, Francisco Guala claims that:

"Problems of internal validity are chronologically and epistemically antecedent to problems of external validity: it does not make much sense to ask whether a result is valid outside the experimental circumstances unless we are confident that it does therein" (Guala, 2003, 1198).

The claim about temporal priority is that we first make inferences about the local environment under study before making inferences about the surrounding world. The claim about epistemic priority is that we come to know the local environment before we come to know the surrounding world.

In the following we problematize the relation between external and internal validity. Our claim is that the two types of validity are deeply intertwined. However, we are not going to attempt to argue for the full claim. We argue only in favour of the part of the claim that is in conflict with the idea behind the internal/external distinction. The argument is directed at showing that internal validity *understood as prior to external validity* has, at least, three epistemologically problematic aspects: experimental artefacts, the implications of causal relations, and how the mechanism is measured. We exemplify the problems associated with experimental artefacts and mechanism measurement by cases from experimental psychology. Each aspect demonstrates how important external validity is for the internal validity of the experimental result.

We end the paper by presenting a different kind of test. Lee Cronbach claims that internal validity, as interpreted by the later Campbell, is a rather meaningless feature of scientific results. If we are right, a Cronbachian attack on internal validity in general must also be mistaken. Since on our understanding internal and external validity are intertwined a successful attack on internal validity would threaten to have adverse effects on external validity. To be consistent with our standpoint the particular conception Cronbach attacks should pinpoint other features than the concept of internal validity has traditionally been assumed to capture.

2) What is internal and external validity?

It is impossible to evaluate whether the perceived tension and the claimed priority of internal validity are justified unless we know more precisely what it is that we make internally valid inferences about and what this validity is supposed to consist in. Below we present three formulations of internal and external validity:

Campbell's early conception: "First, and as a basic minimum, is what can be called *internal validity*: did in fact the experimental stimulus make some significant difference in

this specific instance? The second criterion is that of *external validity*, *representativeness*, or *generalizability*: to what populations, settings, and variables can this effect be generalized?" (Campbell 1957, 297).

Guala's recent conception: "*Internal* validity is achieved when the structure and behavior of a laboratory system (its main causal factors, the ways they interact, and the phenomena they bring about) have been properly understood by the experimenter. For example: the result of an experiment E is internally valid if the experimenter attributes the production of an effect B to a factor (or set of factors) A, and A really is the (or a) cause of B in E. Furthermore, it is externally valid if A causes B not only in E, but also in a set of other circumstances of interest, F, G, H, etc." (Guala 2003, 1198).

Campbell's later conception: "In the new contrast, external […] validity involve[s] theory. Local molar causal validity [, i.e. internal validity,] does not. While this contrast is weakened in the principle of proximal similarity [i.e. external validity], I still want to retain it. The principle of proximal similarity is normally (and it should be) implemented on the basis of expert intuition. […] Our intuitive expectations about what dimensions are relevant are theory-like, even if they are not formally theoretical. Moreover, clinical experience, prior experimental results, and formal theory are very appropriate guides for efforts to make the exploration of the bounds of generalizability more systematic." (Campbell 1986, 76)

Campbell's early conception and Guala's conception show similarity in how they understand external validity. It is about how to generalize what has been found internally. Campbell's later conception differs from both in that the connection between local causal claims and general claims is weakened. The word "local" emphasizes that the claimed validity is limited to "the context of particular treatments, outcomes, times, settings, and persons studied" (Shadish et al. 2002, 54). Local causal claims are "molar" as well. Campbell exemplifies it in the following way: "For the applied scientist, local molar causal validity is a first crucial issue and the starting point for the other validity questions. For example, did this complex treatment package make a real difference in this unique application at this particular place and time?" (Campbell 1986, 69). There is no guarantee that molar claims refer directly to a potential cause. A true molar claim entails merely that

something in the complex it captures is a cause. The difference between Campbell's later conception and Guala's conception is considerable in that respect. Guala's internal validity requires that we understand the causal mechanism that operates in the local case. The later Campbell explicitly opposes such a view as generally true of internal validity. Applied scientists also need internal validity, but they can normally not analyse causation with such precision; "to stay with our problems, we must use techniques that, while improving the validity of our research, nonetheless provide less clarity of causal inference than would a retreat to narrowly specified variables under laboratory control" (Campbell 1986, 70-71). The difference between Campbell's earlier and later understanding of internal validity seems to be one of emphasis primarily. However, the difference between their views of external validity is more significant. External validity is not in general established through representative sampling, and it is not a matter of simple inductive generalisation. First, a cause has to be extracted from the molar situation and then the causal relation is exported to proximally similar cases.

For each of these conceptions there are epistemologically problematic aspects of internal validity. We will focus on three: experimental artefacts, the implications of causal relations, and the measurement of mechanism.

3) Epistemology—the problem of experimental artefacts

Can there be such a thing as an internally valid inference? That clearly depends on whether the methods we use guarantee that we see clearly, i.e. that what we see in the local environment is not in fact an artefact of something else. But some well-known "internally valid" results have in fact been generated by, for instance, the method of randomization or measurement used.

3a) Overconfidence—experimental artefacts

Overconfidence is a psychological phenomenon that refers to an overrating of the correctness of one's judgements. Typically, participants are asked knowledge questions such as "Which city has more inhabitants? Hyderabad or Islamabad?" and are asked to rate how confident they are that their answer on this particular question is correct on a scale

from 50% to 100%. Overconfidence occurs when the mean subjective probability assigned to how correct responses are is higher than the proportion of correct answers. In contrast a participant is calibrated if: "…over the long run, for all propositions assigned a given probability, the proportion that is true equals the probability assigned" (Lichtenstein, Fischhoff and Philips, 1982).

The overconfidence effect can, however, be made to disappear under certain experimental conditions. Some authors (e.g., Gigerenzer, Hoffrage and Kleinbölting, 1991; Juslin, 1994) have claimed that the overconfidence effect is simply an effect of unrepresentative sampling. The basic idea behind the critique is that participants need a certain amount of information in order to make a correct estimate of their performance on a task. When this is not available, they will instead draw on their more general knowledge of the area. If I have no clear intuition on whether Islamabad or Hyderabad is the biggest city in the question above, I might use the knowledge I have of my general competency in geography or what I know about the capitals of Asian countries to produce a confidence judgement. That means that if the knowledge questions are sampled in a skewed way so that they contain more difficult questions than are normally encountered, participants will exhibit overconfidence (i.e. miscalibration). If the knowledge questions posed are instead randomly sampled from representative environments, the overconfidence effect disappears (Gigerenzer et al., 1991; Juslin, 1994).

The early experiments investigating overconfidence were clearly internally valid in the sense that results were robust: The experimental stimuli produced judgments that had the properties of overconfidence. However, they appear to be experimental artefacts, and slight variations in the experimental set up will change the results. There are, however, even more serious allegations against overconfidence – allegations that are especially interesting in this context. In a second set of critique against overconfidence authors such as Ido Erev (Erev, Wallsten and Budescu, 1994) and Peter Juslin (Juslin, Winman and Olson, 2000) claim that overconfidence (and the related hard-easy effect which we will not discuss here) is a product of regression towards the mean. Overconfidence occurs because a participant responding to a difficult task (as the one described above) is more likely to overestimate correctness than underestimating it. In the extreme, a participant that responds at a chance

level cannot be underconfident given the scale 50% to 100% certain that the response is correct. This explains also why the representatively sampled knowledge questions (of intermediate difficulty) made the overconfidence effect disappear. The artefact is not produced by the knowledge questions as such, but depend rather on features inherent in the experimental situation: it is difficult to conceptualize a scale measuring certainty that would not have endpoints such as these.

4) Epistemology—the problem of causation

Whether there can be an internally valid inference also depends on the nature of what is inferred to. Normally, as we have seen in 2) the inference is causal. Now, there are many concepts of causation. Some of these are clearly of a kind that does not support inferences that are primarily internally. For instance, someone operating with a notion of causation similar to one of those that Kant, Hume, or Mill relied on will judge internally valid inferences to causal matters impossible. For each of those causal concepts the implications of causation, regardless of whether it has to do with the notion of sufficiency or necessity, go beyond the local environment. If there is a causal relation in the local environment it follows that this holds also outside this environment. And, trivially, it holds that if it does not hold outside the environment it cannot hold inside either. Hence such concepts of causation warrant neither the alleged temporal nor epistemological priority of internal validity.

It is in fact a long distance between traditional causal concepts and causation that is suitable for being primarily internally validly inferred to. However, more than one advocate of randomised controlled trials adopts a view on which an intervention study underwrites a positive causal inference. Consider the following quote from David Papineau:

"You take a sample of people with the disease. You divide them into two groups at random. You give one group the treatment, withhold it from the other [...] and judge on this basis whether the probability of recovery in the former group is higher. If it is, then T [treatment] must now cause R [recovery], for the randomization will have eliminated the

danger of any confounding factors which might be responsible for a spurious correlation." (Papineau 1994, 439)

This is excessively optimistic for reasons having to do with the possible artefacts of randomization (cf. Shadish et al., 2002, Ch. 2) and the more general points that we have already pressed, but that is, not the present point. Let us assume that randomization is successful in the desired respect. Papineau's modified position seems to rely on a concept of causation given which in the relevant cases causation is entailed by (i.e. is unproblematically inferable from) the fact that the relative frequency of R in the intervention group is higher than it is in the control. Thus, for instance, the concept of cause employed is not that causes are sufficient in the circumstances, nor that they are necessary. This is plainly not so since neither kind of causation is entailed by the experimental fact (cf. Persson 2009).

5) Epistemology—the measurement of mechanism

How mechanisms are measured has a strong impact on the results obtained. As we saw in the case of overconfidence the choice of measurements can have unintended side effects, but the relation between how stimuli are presented and the effects that are measured is more complex than so. An interesting example comes from psychophysics and concerns range effects, i.e., effects due to the fact that participants receive more than one experimental condition.

5a) Range effects– the measurement of mechanism

Poulton (1975) presents a number of different range effects demonstrating how the order in which stimuli is presented in itself affect the result, or the type of mechanism that is being observed (an "unbiased" perceptual judgment, or judgments mediated by range effects – in themselves mechanisms). We will use the simplest example, where the range in which a stimulus is presented influences how far apart different stimuli are judged to be. In the case of Figure 1 the slope of perceived distances between stimuli is radically different when the end points are L_1 and L_2, rather than S_1 and S_2 when $\varnothing$ represents the physical magnitude and ψ the subjective (perceived) magnitude.

[INSERT FIGURE 1 POULTON; SEE LAST PAGE]
Figure 1. Adapted from Poulton, 1968.

Since participants' pre-conceptions of what the range of stimuli is will affect their responses, the "external validity" of the stimuli (in this context how well the range it introduces, or the range the experimenter assumes, matches participants' pre-conceived range of stimuli) determines whether the results obtained *in the laboratory* correctly capture the features of the mechanism operating there. Hence, in cases like these, external validity is a requirement for internal validity. Note that this potentially false estimate of the function has prefect internal validity. Given the range, the stimuli really do cause the response, and we have a fair grasp of what the mechanisms are.

Poulton himself, however, treats the results differently than we do: "All experimental data are not equally valuable. A theoretical model is unlikely to be better than the data which has shaped it. If data are of restricted validity as a result of unrepresentative sampling or the independent variables or of uncontrolled transfer effects, a model based upon the data is not likely to have great generality. This is the case however much data the model can fit, provided all the data has been generated using the same inadequate techniques of sampling or experimentation" (Poulton 1968, 1). We do not disagree with Poulton, but in contrast to him we emphasize that the core issue here is how internal validity is to be guaranteed unless range effects are properly understood. And this will happen only when extra-experimental factors (such as participants' pre conception of the range that is to be introduced) are properly understood. Thus we would like to maintain that the case of the perceptual mechanisms at the mercy of range effects internal and external validity cannot be treated as separate entities.

6. The difficulty of adapting systems

A straightforward extension of the above observations about the co-dependence of external and internal validity is to be found in Egon Brunswik's work on representativeness. What he adds to the discussion is a focus on the difficulties in observing an organism that adapts

to the circumstances in which it exists: "The concept inherent in functionalism that psychology is the science dealing with the adjustment of organisms to the environment in which they actually live suggest the need of testing any obtained stimulus-response relationship in such a way that the habitat of the individual, group, or species is represented with all of its variables, and that the specific values of these variables are kept in accordance with the frequencies in which they actually happen to be distributed." (Brunswik, 1944, 69).

Note, however, that here the focus is exclusively on the adaptive character of human cognition (in Brunswik's case the perceptual system). If the aim of an experiment in psychology is to understand the functioning of different psychological mechanisms (in the form of stimulus-response relations), then the quality of this finding is just as dependent on whether the psychological mechanism has been properly activated as it is on whether the results can be replicated. This is not only a question about how the result will generalize to other settings (external validity) – it is a question about whether a proper result has at all been generated (internal validity). Thus, for psychological mechanisms that can be assumed to have an adaptive character, external validity (or certain aspects of it) appears to be prior to internal validity: It is more important that an experiment measures what it aims to measure than that the result internally valid.

6a Is the study object human cognition or the environment?

Egon Brunswik is one of the psychologists that have most clearly advanced the idea that external validity has to be taken into account if we are to understand the human mind at all. In his own words: "psychology has forgotten that it is a science of organism-environment relationships, and has become a science of the organism" (Brunswik, 1957, 6). His remedy to this difficulty was the notion of representative design (Brunswik, 1955), and, in particular, his use of representative sampling while studying perceptual constants (Brunswik 1944).

In his 1944 study, Brunswik wanted to understand whether the retinal size of an object could be used to predict its actual size. In order to establish the relationship between retinal size and object size, participants were followed for several weeks and stopped at random

intervals. For whatever object they were looking at, at that point, retinal size, object size, and distance were measured. Since the objects taken into account were the objects actually attended to by participants in their daily environments, Brunswik could estimate the real-life predictive power of retinal size for object size. His conclusion was that the retinal size had some predictive power regardless of the distance to the object.

Note that Brunswik's method as described here is *only* a method for understanding the environment. In order to explain how participants judge the size of objects, it has to be combined with a demonstration that retinal size is used to predict object size. However, the controlled experiment that can be used to test this hypothesis will not help us understand how predictive retinal size is of object size. This requires a method such as Brunswik's.

Note also that the method of representative sampling is only possible in so far as the researcher *already* has a clear understanding of the cognitive process under investigation. Unless we have some idea of which aspects of the environment are accessed by the cognitive mechanism, methodological shortcuts such as representative sampling are not possible. Simply stated, we have to know what to measure in order to measure it, also when the measurement is done through random sampling. Campbell, of course, notes this problematic issue in the context of random sampling of *participants* (note the difference in emphasis). He points out that: "… the validity of generalizations to other persons, settings, and future (or past) times would be a function of the validity of the theory involved, plus the accuracy of the theory-relevant knowledge of the persons, settings, and future periods to which one wanted to generalize […]. This perspective has already moved us far from the widespread concept that one can solve generalizability problems by representative sampling from a universe specified in advance" (Campbell 1986, 71).

Also other methodologically inclined psychologists have reflected upon the co dependency of the environment and the agent. Often this is conceptualized as the difficulty of understanding whether what is being observed is a feature of the participant's internal processing or a feature of the task environment. Thus Ward Edwards (1971) observes that: "My own guess is that most successful models now available [in psychology] are successful exactly because of their success in describing tasks, not people …modelling tasks is different from modelling people, [we need] to hunt for tools for modelling tasks,

and to provide linkages between models of tasks and models of people". And this difficulty has it roots in precisely the difficulty of making controlled experiments that observe features of a cognitive system designed for adapting to the circumstances. Or in Campbell's own words: "Both criteria [external and internal validity] are obviously important although it turns out that they are to some extent incompatible, in that the controls required for internal validity often tend to jeopardize representativeness" (Campbell 1957, 297).

7) Cronbach's challenge

Let us now set the objections against the possibility of internally valid inferences aside. Let us grant that the problems of randomization, measurement and causation can be dissolved by appropriate adaptive measures. Even so the question whether internal validity should be given priority remains:

"I consider it pointless to speak of causes when all that can be validly meant by reference to a cause in a particular instance is that, on one trial of a partially specified manipulation t under conditions A, B, and C, along with other conditions not named, phenomenon P was observed. To introduce the word cause seems pointless. Campbell's writings make internal validity a property of trivial, past-tense, and local statements." (Cronbach 1982, 137)

Cronbach's point translates nicely to what we have argued here. To the extent that there is a variety of causation that can be fully examined in such a way that it underwrites a positive causal inference—for instance, by a randomized controlled trial—then that variety of causation is not very scientifically valuable. What should we do with these past tense, local statements concerning highly artificial experimental contexts? They seem trivial as scientific results. The only way this kind of trivial causal statements could prove useful is if they connect with more substantial ones. In other words, internal validity of this kind could have a value in relation to external validity as providing one of the instances externally valid claims have to be true about. Now, internal validity is not prior to external validity in any interesting sense. If anything, it seems secondary. It should be noted that Campbell (1986, 70) acknowledges this: "The theories and hunches used by those who put

the therapeutic package together must, of course, be regarded as corroborated, however tentatively, if there is an effect of local, molar validity in the expected direction".
However, this relationship between internal and external validity is important. Cronbach's challenge might be reconstructed as a counter argument to our claim that internal and external validity are intertwined. It might be constructed as the view that internal validity is redundant. As we have seen our response is: 1) to the extent that the causation internal validity concerns is substantial, external validity is needed as part of the evidence; 2) to the extent that the causation is of a trivial form, this kind of causation might still be important as one of the instances that is needed to prove external validity. (There is, of course, a third possibility as well, that all genuine causation is local.)

8) Priorities reconsidered

However critical we have been of attempts to prioritize internal validity, there is a last argument that can be made in its favour, and it is elegantly (and fittingly) made by Campbell in the following passage: "If one is in a situation where either internal validity or representativeness must be sacrificed, which should it be? The answer is clear. Internal validity is the prior and indispensable consideration. The optimal design is, of course, one having both internal and external validity. Insofar as such settings are available, they should be exploited, without embarrassment from the apparent opportunistic warping of the content of studies by the availability of laboratory techniques. In this sense, a science is as opportunistic as a bacteria culture and grows only where growth is possible. One basic necessity for such growth is the machinery for selecting among alternative hypotheses, no matter how limited those hypotheses may have to be." (Campbell 1957, 310). Although we do not believe that internal and external validity can be treated separately – or even chosen between in the way suggested by Campbell – we fully agree that scientific research will have to take whatever routes are available.

References

Brunswik, E. (1944). Distal focussing of perception: size constancy in a representative sample of situations. *Psychological Monographs, 56*(1), Whole No.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review* 62(3): 193 - 217.

Brunswik, E. (1957). Scope and aspects of the cognitive problem. In J.S. Bruner, E. Brunswik, L. Festinger, F. Heider, K. F. Muenzinger, C. E. Osgood and D. Rappaport (eds.). *Contemporary approaches to cognition*. Cambridge: Harvard University Press.

Campbell, D., T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54 (4): 297-312.

Campbell, D., T. (1986). Relabeling internal and external validity for applied social sciences. In W., M., K. Trochim (ed.). *Advances in Quasi-Experimental Design and Analysis. New Directions for Program Evaluation*, no 31. San Francisco: Jossey-Bass, Fall 1986.

Cronbach, L., J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass Publishers.

Edwards, W. (1971). Bayesian and regression models of human information processing – A myopic perspective. *Organizational Behavior and Human Performance*, 6: 639-648.

Gigerenzer, G., Hoffrage, U. and Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review* 98(4): 506-528.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science*, 70(5): 1195-1205.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of items. *Organizational Behavior and Human Decision Processes* 57: 226-246.

Juslin, P., Winman, A., and Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect. *Psychological Review* 107(2): 384-396.

Lichtenstein, S., Fischhoff, B. and Philips, L., D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman and A. Tversky (eds.). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

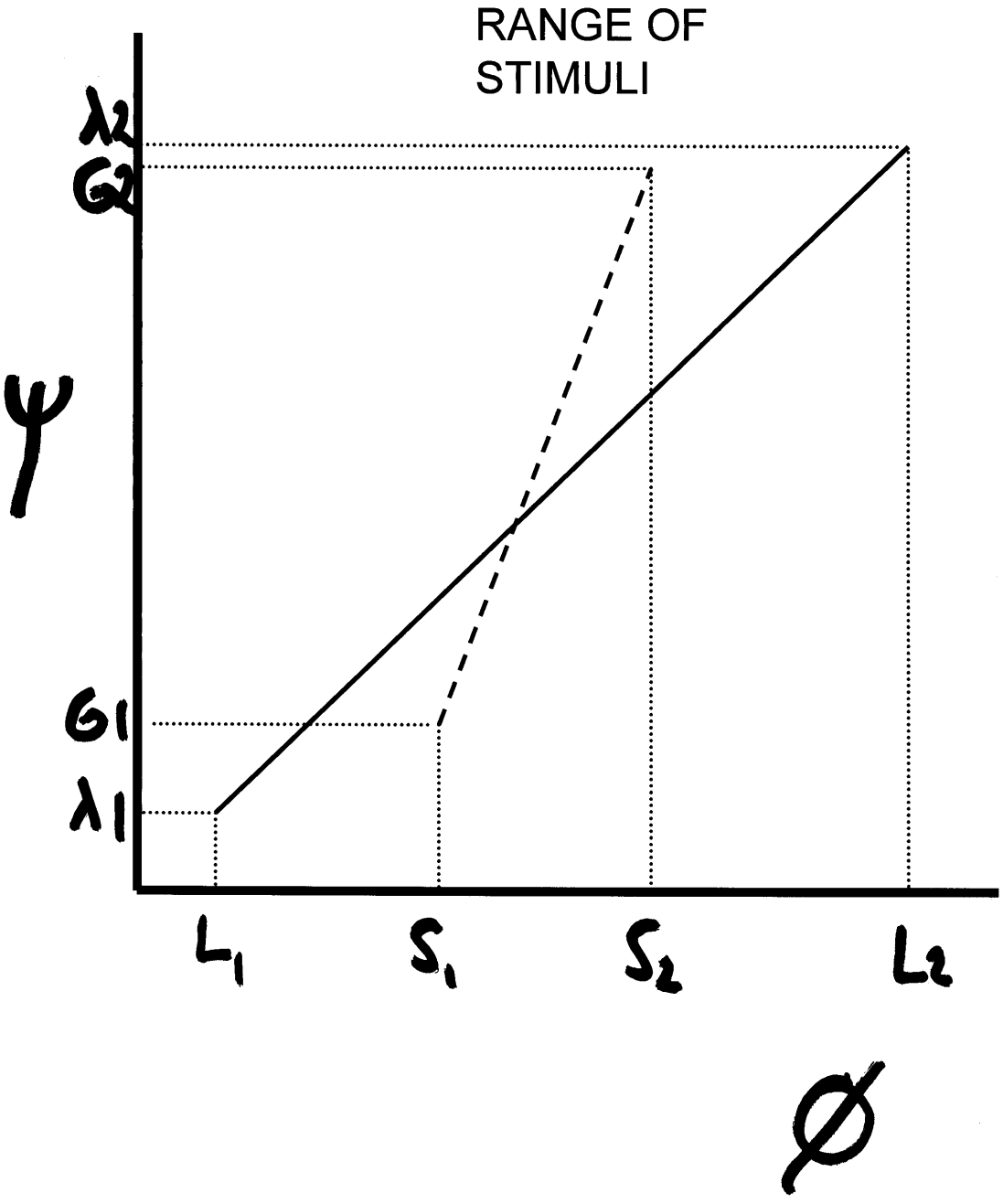
Papineau, D. (1994). The virtues of randomization. *British Journal for the Philosophy of Science*, 45(2), 437–450.

Persson, J. (2009). Semmelweis's methodology from the modern stand-point: intervention studies and causal ontology. *Studies in History and Philosophy of Biological and Biomedical Sciences* 40: 204–209

Poulton, E., C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, 69: 1-19.

Poulton, E., C. (1975). Range effects in experiments on people. *The American Journal of Psychology*, 88(1): 3-32.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.


RANGE OF
STIMULI
λ2
G2
ψ
G1
λ1
L1
S1
S2
L2
ø

Physical Symmetries, Overarching Symmetries, and Consistency

Abstract

In this paper I provide an account of physical symmetries, which are defined relative to a specific physical theory, and overarching symmetries, which hold across many different physical theories. I outline two general strategies for uniting disparate physical symmetries under the same overarching symmetry, calling the first “realist” and the second “conventionalist”. Finally, I argue that, should physicists and philosophers be interested in finding symmetries that do interesting and helpful physical and metaphysical work, the realist strategy would serve them better than the conventionalist strategy.

1. Introduction

Symmetries play an important role in our understanding of our best physical theories, giving us important information both about the world and the physical theories we posit to aptly represent the world. The project of this paper is to provide an account of two kinds of symmetries and justify a particular approach to investigating these symmetries. I divide symmetries into two classes, namely "physical" symmetries defined within the context of a specific physical theory and "overarching" symmetries defined across numerous physical theories. My first task in this paper is to provide an account of physical symmetries; however, most of this paper will focus on overarching symmetries and the relationship between overarching and physical symmetries. In particular, I will discuss "realism" and "conventionalism", two different stances towards certain overarching symmetries, and provide several examples of how overarching symmetries may guide scientific inquiry. My central claim is that should we wish to use overarching symmetries as guides to what theories we should accept as the best successors to rejected theories, guides to the underlying metaphysical structure of the world, or indicators of inter-theoretic incompatibilities, we ought to take overarching symmetries to act "realist"-ically across physical theories. I thus argue that both realists and conventionalists have good reason to care about overarching symmetries that act consistently across different physical theories.

2. What are Physical Symmetries?

Symmetries are of interest because they reveal meaningful operations under which basic structures are preserved. For instance, a starfish's radial symmetry tells us that the starfish's shape is preserved under rotations of certain intervals. And so in fundamental physics as well, what

counts as a symmetry is a kind of operation that leaves a certain structure invariant[1]. There are essentially three features one must identify in order to understand any symmetry: the objects acted on by the symmetry, the way in which the symmetry transforms those objects, and the relevant structure(s) left invariant by the symmetry. Theses three features are invariably linked. Consider spacetime symmetries in special relativity, for instance. Such symmetries are maps from spacetime points to spacetime points that leave the Minkowski spacetime interval invariant. These symmetries are thus maps from points defined on a particular topological space (spacetime) to points on that same particular topological space that leave some particular feature of those points (the spacetime interval between points) invariant. In order to better understand theory-relative physical symmetries, then, we should investigate the sorts of topological spaces and invariant structures utilized by all physical theories.

Physical theories, as I use the term for the purposes of this paper[2], are essentially ordered tuples of the form <E,X,U,N>, where E is the set of equations utilized by a particular theory, X is the set of all independent variables appearing in the equations in E,[3] U is the set of all dependent variables appearing in the equations in E, and N is the "interpretation function" of that physical

[1] Something like this general notion of symmetry can be found in, for instance, Belot (2011), Brading and Castellani (2007), Healey (2009), Ismael and van Fraassen (2003), and Roberts (2008).

[2] I am not trying to give a full account of physical theories here; rather, what I refer to as a "physical theory" is a set of necessary features that I take all physical theories to have, all of which I will exploit shortly.

[3] Over the course of this paper I will refer to X as both the set of independent variables and the space whose axes correspond to the independent variables in X. Context will determine which of the two I mean in each case, and I will similarly equivocate with my usage of U.

theory (more on that shortly). By dependent variables here I mean variables that can be represented as functions of the independent variables and whose derivatives in terms of these independent variables we take to be of physical interest. For instance, in Newtonian mechanics the variable representing the position of a ball rolling down an inclined plane can be represented as a function of the time that has passed since the ball was released, and so the variable "position" here is taken to be a dependent variable u and "time" is taken to be an independent variable t. There may be some conventional element in selecting which variables are dependent and which are independent for any particular physical theory; for instance, because the position u can be given in terms of the time t by $u=f(t)$, it is also the case that, for invertible f , $t=f^{-1}(u)$, so we may free to treat position or time here as the dependent variable. Alternatively, we may take the velocity of the ball, $v=dx/dt$, to be more fundamental than its inverse dt/dx and so take there to be a principled reason for treating t as the independent variable here instead of x. Typically, the variable that is easier for experimenters to control or manipulate is taken to be the independent variable, but this need not always be the case. My point here is only that it is up to the theory we are interested in to tell us what quantities to care about and whether to treat them as independent or dependent.

The equations in E[4] are conditions on functions of independent variables, dependent variables, and, in the case of differential equations, derivatives of the dependent variables with respect to the independent variables, that require these functions be zero[5]. So, for instance, in cases where there is only one independent variable x and one dependent variable u, an algebraic

[4] This account follows the one given by Olver (1993).

[5] Though everything I say in this paper can be extended to an account that takes E to include differential as well as algebraic equations, due to length considerations, I will restrict my analysis in this paper to theories whose equations are all algebraic.

equation can be represented by the condition that the function e(x,u)=0. An equation can thus be characterized by the function it sets equal to zero, so I will take the elements of the set E to be the functions that the equations of the physical theory set equal to zero. Equations allow us to represent physical constraints rather than mere mathematical peculiarities because we take the variables appealed to by our physical theories to represent measurable features of the physical world, and so physical theories necessarily contain what I call an interpretation function N, which is a map from the variables appealed to by a physical theory and functions of these variables to measurement processes in the physical world that provide the values for these variables. Interpretations are necessary components of physical theories because two physical theories may utilize similar equations and variables to different ends. For instance, the heat equation:

$$d(x,t,u)=\partial u/\partial t-\partial^2 u/\partial x^2=0$$

could represent one-dimensional particle diffusion or the one-dimensional temperature change in, say, a metal bar. Since we measure position probability density functions (the "u" appealed to in the differential equation in the first case) and temperature (the "u" appealed to in the second case) differently, we can rely on N to distinguish these two different applications of the same differential equation[6].

[6] One may object here that physical theories do not all come equipped with particular interpretations; after all, many physical theories (most notably quantum mechanics) have numerous interpretational difficulties, and so stipulating that a physical theory like quantum mechanics comes with "an interpretation" seems strange. Note, however, that my technical usage of the "interpretation function" differs from what is typically meant by an interpretation of a physical theory: the job of my "interpretations" is only to tell us what measurement processes give us the values for the variables postulated by a physical theory, not to tell us anything about

Let's turn from my discussion of what physical theories are to the question of what topological space the objects of the physical theory live in. Algebraic equations, which are functions of independent and dependent variables only, can be represented by n-1-dimensional submanifolds of the n-dimensional space of dependent and independent variables X and U, which I will henceforth call the *variable space* X×U. Each point in this space corresponds to a set of values for the variables in X∪U, and the theory differentiates between those sets of values that are *physically possible* according to a particular equation, which correspond to points in the equation's submanifold, from those sets of values which are not physically possible according to that equation. Few theories contain singleton sets of equations, and so we are typically interested not in each submanifold on its own but rather the intersections of the submanifolds corresponding to all of the equations in a physical theory's E. It is these submanifolds that determine which sets of variable values the theory takes to be physically possible simpliciter and which values it does not. We can define a *solution* of an algebraic equation as a map from the space of independent variables X to the space of dependent variables U such that, when this map is identified in the obvious way with a subset of the variable space X×U, this subspace lies entirely within the submanifold determined by the equations. This notion of a solution extends naturally to characterize solutions of differential equations as well[7].

Following this definition of a solution, one can now posit a notion of symmetry that lines up nicely with the previously discussed notion of a spacetime symmetry. A symmetry[8] of a

---

the underlying metaphysics of the world described by this theory.

[7] See Olver (1993) for the details.

[8] I will restrict my talk of symmetries in this paper to point symmetries. Those interested in generalized symmetries should consult chapter 5 of Olver (1993).

physical theory T is a transformation that maps points in X×U to other points in X×U that keeps the solution structure the same. That is, a symmetry cannot map the points that lie entirely within the submanifold determined by T 's equations to points outside of the submanifold determined by T's equations. The objects transformed by a physical symmetry are points in a theory's variable space, and the structure preserved by a physical symmetry transformation is the solution structure of that space. The set of all transformations of a physical theory, along with a binary operation that defines the product of any two such transformations, constitutes a group.

The search for physical symmetries is thus just the search for the group of symmetries that leaves the solution structure of a variable space invariant. I will forego some technical details here, but there are available mathematical methods that take advantage of the infinitesimal generators of groups that allow us to determine what the symmetry group of any particular function is[9]. So, in short, once we restrict our attention to physical theories that utilize only algebraic equations and define physical symmetries as I have, there are mathematical results that make the calculation of these symmetries (relatively) easy in many cases.

3. The Problem of Overarching Symmetries

My account of physical symmetries in the last section explains what it is to be a symmetry of a physical theory, but there are plenty of contexts in which one may speak of a symmetry without referring to a particular physical theory. For instance, physicists may speak of translations, rotations, parity reversal, and the consequences of these symmetries without referring to any particular theory. Call such symmetries “overarching symmetries”. No sooner have we defined these symmetries, however, than we are faced with the following problem:

[9] Those who would like more detail on how this process works should consult Section 2.1 Olver (1993).

what, exactly, is the relationship between theory-relative symmetries (physical symmetries) and theory-independent symmetries (overarching symmetries)? When should we count two physical symmetries as corresponding to the same overarching symmetry? Or, put more formally: suppose that some theory $T_1=<E_1,X_1,U_1,N_1>$ with n independent variables and m dependent variables is invariant under the point symmetry transformation S_1, which operates on a point $p=(x^1_1,\ldots,x^n_1,u^1_1,...,u^m_1)$, where $x^i_1 \in X_1$ and $u^i_1 \in U_1$, as follows: $S_1(p)=(f_1(x^1_1,\ldots,x^n_1,u^1_1,\ldots,u^m_1), \ldots,f_{n+m}(x^1_1,\ldots,x^n_1,u^1_1,\ldots,u^m_1))$ for some functions $f_1,\ldots,f_{n+m}$. Now, suppose there is another theory $T_2=<E_2,X_2,U_2,N_2>$ with j independent variables and k dependent variables which is invariant under the point symmetry transformation S_2, which operates on a point $q=(x^1_2,\ldots,x^j_2,u^1_2,\ldots,u^k_2)$, where $x^i_2 \in X_2$ and $u^i_2 \in U_2$, as follows: $S_2(q)=(h_1(x^1_2,\ldots,x^j_2,u^1_2,\ldots,u^k_2),\ldots,h_{j+k}(x^1_2,\ldots,x^j_2,u^1_2,\ldots,u^k_2))$ for some functions $h_1,\ldots,h_{j+k}$. Under what conditions can we say that T_1 and T_2 are invariant under the same overarching symmetry transformation?[10].

Another complicating feature is the fact that variables that appear in two different physical theories may refer to the same feature of the physical world or may be calculated by exactly the same methods. For instance, we may refer to the length of a metal bar in the context of both classical thermodynamics and special relativity. In both cases the length of the bar can be

[10] I will leave aside for now the question of whether there is any *natural* property à la Lewis (1983) picked out by all overarching symmetries. My account allows that the question "Is X really an overarching symmetry?" may or may not be substantive; the real project for those investigating overarching symmetries, I take it, isn't to figure out what overarching symmetries correspond to the *real* overarching symmetries in the world (if there are any) but rather to lay down a useful criterion for what constitutes a particular overarching symmetry and argue why this criterion is the most useful one.

calculated by the same measurements because the physical quantity, length, is the same. So, we may be especially worried about how to pick out the symmetry corresponding to S_2 in the formalism above when, for some a in $X_1 \cup U_1$ and b in $X_2 \cup U_2$, $N_1(a)=N_2(b)$ or $N_1(f(a))=N_2(f(b))$ for some f. Call any two such variables *empirically identical*.

Though I cannot address all of the criteria one could lay down as good candidates for determining which physical symmetries fall under the same overarching symmetry, I can outline two reasonable positions regarding these criteria. The difference between these two positions rests on how strong a criterion one utilizes for determining when physical symmetries are related to one another by an overarching symmetry and when they are not. For our purposes, I propose (but will not defend at length) a criterion based on the following definition, though any equally strong criterion should do for the points I will make in the rest of the paper:

> (Consistency): Take $T_1=<E_1,X_1,U_1,N_1>$ to be a physical theory invariant under the symmetry transformation S_1, which maps each coordinate v^i_1 in $X_1 \cup U_1$ to $f_{i1}(v^1_1, v^2_1,\ldots)$, and take $T_2=<E_2,X_2,U_2,N_2>$ to be a physical theory invariant under the symmetry transformation S_2, which maps each coordinate v^j_2 in $X_2 \cup U_2$ to $f_{j2}(v^1_2,v^2_2,\ldots)$. There are two sets of variables $A \supseteq X_1 \cup U_1$ and $B \supseteq X_2 \cup U_2$ such that $\forall v^i_1 \in X_1 \cup U_1$, if $\exists v^j_2 \in X_2 \cup U_2$ such that v^i_1 and v^j_2 are empirically identical, then $v^i_1 \in A$, and $\forall v^j_2 \in X_2 \cup U_2$, if $\exists v^i_1 \in X_1 \cup U_1$ such that v^j_2 and v^i_1 are empirically identical, then $v^j_2 \in B$. S_1 and S_2 are *consistent* if and only if, $\forall v^i_1 \in A$ and $\forall v^j_2 \in B$ for non-empty A and B, the empirical identity of v^i_1 and v^j_2 implies $f_{i1}=f_{j2}$ for some fixed values of all the variables $v^n_1 \notin A$ and $v^m_2 \notin B$ for which f_{n1} and f_{m2} are not constant functions.

Despite its formal complexity, Consistency is intuitively easy to understand. Essentially, two

physical symmetries are consistent only if both symmetries treat the "same variables" in the "same way". By "same variables" here I mean variables that are empirically identical, and by treating these variables in the "same way", I mean that, S_1 and S_2 use the same functions to transform empirically identical variables, ignoring changes involving the empirically non-identical variables. To give an example: suppose that one theory has a symmetry S_1 that transforms points in its variable space as follows: $S_1(x,y,z)=(z(x+y),y,z)$, and suppose that another theory has a symmetry S_2 that transforms its points as follows: $S_2(x,y,w)=(x+y+w,y,w)$, where I have used the same variable name to denote variables in different theories that are empirically identical to one another. S_1 and S_2, according to my definition, are consistent since $S_1(x,y,1)=(x+y,y,1)$ and $S_2(x,y,0)=(x+y,y,0)$; however, S_3, which acts as follows: $S_3(x,y,w)=(x+y,x+y,w)$, is not consistent with S_1 since y!=x+y.

We can now define at least two positions on overarching symmetries, the first of which requires that all physical symmetries associated with some overarching symmetry be consistent[11]. I will call such a position "realist". The second position denies consistency as a constraint on overarching symmetries in favor of some weaker criterion. I will call such a position "conventionalist"[12]. The realist holds, basically, that an overarching symmetry unites consistent physical symmetries across physical theories. For instance, if T_1 has a coordinate x that refers to time and T_2 has a coordinate t that refers to time, then if T_1 is invariant under "time translation"

[11] Or as I said before, some equally strong or stronger criterion for overarching symmetry-hood. For the rest of the paper, interpret all appeals to consistency as appeals to "some criterion at least as a strong as consistency".

[12] Note that, as defined, realism and conventionalism are relative to the particular overarching symmetry under consideration. There is no reason to disallow, say, simultaneous realism about time reversal and conventionalism about gauge transformations.

via a physical symmetry that takes x to x+a, then T_2 can only be invariant under the same overarching symmetry of "time translation" if it is invariant under a transformation that takes t to t+a (modulo some variables that may appear in one theory but not the other). More importantly, if two theories refer to some common set of parameters (e.g. time, position, and momentum), then the two theories are invariant under the same overarching symmetry, according to the realist, only if there is some symmetry transformation defined on the first theory that treats these parameters the same way that some symmetry transformation defined on the second theory does.

Conventionalists, on the other hand, embrace some weaker constraint, making it easier for two different symmetries to be identified with the same overarching symmetry. For instance, in the previous example, one sort of conventionalist could argue that, despite the fact that T_1 is only invariant under a symmetry that takes x to x+a and the T_2 is only invariant under a symmetry that takes t to –t+a, T_1 and T_2 may still be invariant under the same overarching symmetry because both symmetries satisfy the weaker criterion of transforming empirically identical variables by some procedure that adds an arbitrary constant. Some conventionalist criteria may come quite close to realism's. For instance, the conventionalist may adopt a criterion that overarching symmetries link up physical symmetries that act consistently on one particular variable. Regardless of the specific criterion, however, what separates realists and conventionalists is that conventionalists allow a strictly larger set of physical symmetries to qualify as potential instances of a particular overarching symmetry than realists allow.

The difference between realism and conventionalism clearly becomes salient when trying to determine certain important features of physical theories. We may be interested, as many philosophers of time and philosophers of physics have been, in the question of whether or not the fundamental laws of physics are invariant under time reversal. Realists trying to answer this

question will look at each fundamental physical theory to determine whether or not its time reversal symmetry (if it exists) can be consistently unified with the time reversal symmetries of the other fundamental physical theories. Conventionalists, meanwhile, will likewise examine each physical theory to be sure that there is some point symmetry suitably called "time reversal" (under some weaker standard) under which it is invariant, and should they find at least one transformation for all fundamental physical theories, they will be happy to agree that the fundamental laws of physics are invariant under time reversal. Henceforth I will refer to the general strategy of looking for a consistent overarching symmetry as the realist strategy and the strategy of searching for physical symmetries that satisfy some weaker criterion on overarching symmetries as the conventionalist strategy, though I should point out that conventionalists can adopt the realist strategy too. In the remainder of this paper, I will argue that the realist strategy ought to be the first one that philosophers and physicists pursue when investigating overarching symmetries by suggesting some philosophical work that consistent overarching symmetries can do for us that inconsistent overarching symmetries cannot.

4. Overarching Symmetries, Theory Change, and Ontology

Overarching symmetries help us to determine which new theories should replace older, falsified theories. Typically, when an earlier theory has been rejected and physicists are searching for a new theory to replace it, physicists don't just start from scratch. They assume that many of the features of the old theory were, in fact, correct, assuming the theory in question had a history of empirical successes. The old theory must have gotten *something* right to be successful, so since new theories should be strictly better than the theories they replace, we need a way to identify those successful features of our old theories and carry those features over into

our new theories. Successful features of older theories can thus constrain what good candidates for these theories' replacements should look like.

Symmetries are typically features of the world that carry over from older theories to newer theories. Imagine, for instance, a fictitious history of physics in which Newtonian mechanics is rejected after observing the behavior of particles traveling close to the speed of light. Under such circumstances, we would be justified in searching for a successor to Newtonian mechanics like special relativity that isn't invariant under Galilean boosts; however, we have no reason to reject the symmetries of spatial and temporal translation under which Newtonian mechanics was invariant. So even if our evidence gives us good reason to reject important features (including symmetries) of a previous theory, that same evidence may still lead us to uphold some symmetries of the old theory.

Realist overarching symmetries are the most likely to yield useful results when trying to carry symmetries over from failed theories to their successors. Realist symmetries guarantee that the new theory's overarching symmetries differ as little from their predecessors as possible, as least as far as symmetries are concerned. This is not the case with conventionalist symmetries. Take theory A, which is invariant under the symmetry $S_1(x,y,z)=(z(x+y),y,z)$ and which is to be replaced by either theory B, which is invariant under the symmetry $S_2(x,y,w)=(x+y,y,w)$, or theory C, which is invariant under the symmetry $S_3(x,y,w)=(x+y,x+y,w)$. The realist singles out B as the best successor to A since they share an overarching symmetry while the conventionalist may take both B and C to be equally good successors to A since they share an overarching symmetry even though intuitively S_1 seems much closer to S_2 than S_3. If we want symmetries to help us choose useful successor theories, then it seems best to adopt the realist strategy.

Overarching symmetries may also serve as guides to extracting metaphysics from our

best available theories. In some cases, especially when the symmetries in question are spacetime symmetries, the failure of a particular theory to be invariant under an overarching symmetry provides us with a good reason to think that there is some special structure in the world whose existence keeps the theory in question from being invariant under that overarching symmetry. Those who adopt Earman's (1989) symmetry principles, for instance, maybe take the asymmetries of dynamical laws to suggest geometrical features of spacetime.

But symmetries also help to pare down metaphysical commitments; for instance, according to a view I will call physical equivalence (PE) embraced by Baker (2010), Ismael and van Fraassen (2003), and North (2009), symmetries act as a guide to surplus structure. PE holds that the notational differences between two solutions related by a symmetry do not correspond to deep physical differences between two physical states while a notational difference between two solutions which are not related by a symmetry does. Leibniz shift arguments provide an example of PE in action: because our universe is spatial translation invariant, we have good reason to believe that there are no physical structures like, say, absolute space that privilege one position in space over another.

Given these examples, one may feel compelled to adopt either the realist or conventionalist strategy depending on one's projects in metaphysics. If one prefers metaphysical desert landscapes, one may be tempted to adopt both PE and the conventionalist strategy since such an approach will find more solutions to be essentially identical to one another, thus eliminating from our fundamental metaphysics interesting properties that could distinguish one solution from another[13]. Those more interested in using PE to find asymmetries and use them to

[13] It is worth noting, though, that some conventionalist strategies may be so permissive as to undermine the motivation for PE in the first place and may not provide a rich enough metaphysical background to draw the kinds of useful distinctions we want to draw.

posit metaphysical structures, on the other hand, may be more interested in the realist strategy and its stricter criterion on overarching symmetries. At the moment, metaphysical considerations alone don't seem to privilege either account over the other except insofar as one finds one of these projects more interesting than the other.

But the relationship between symmetries and metaphysics does give us another reason to think that, no matter our projects in metaphysics, we should focus our attention on those symmetries that transform the same objects and relations consistently across physical theories. If we do think that symmetries can help us determine the ontological commitments of our physical theories, and if we think that all of the physical theories are, in fact, representations of the same objective world, then the ontologies identified by different physical theories should line up nicely with one another. For instance, if one theory is committed to the existence of a preferred inertial reference frame (as some formulations of quantum mechanics are) and another theory is committed to the physical equivalence of all inertial reference frames (in the way that special relativity is if we assert something like PE for velocity boosts), we have good reason to think that at least one of these theories is wrong since at most one can accurately represent the way the world actually is.

If we want a consistent picture of what the world is like from these theories, and if we think that symmetries can tell us something about the ontological commitments of our physical theories, then even the conventionalist needs to accept the fact that consistent overarching symmetries will be more useful than inconsistent overarching symmetries. Consider the following case: take the two theories T_1 and T_2, both of which are the best currently available physical theories with empirically identical variable x, which is position. T_1 is invariant only under the symmetry $S_1(x,y)=(-x,y)$ while T_2 is invariant only under $S_2(x,z)=(-x,1/z)$. S_1 and S_2 are

consistent, and so they fall under the same realist overarching symmetry. This realist overarching symmetry (let's call it "parity reversal" for short) suggests that there is no fundamental structure that distinguishes between "left" and "right" in the universe since for every solution to the equations in T_1 or T_2 there is a "parity reversed counterpart" related to this solution by S_1 or S_2 respectively that is also a solution to the equations in T_1 or T_2. In short, there is no dynamical structure that treats objects on the "left" differently from objects on the "right", and so, since all of our best physical theories are invariant under this parity reversal symmetry, we have good reason to think that there is no fundamental feature of our universe that distinguishes between left and right. Since realists utilize a stricter overarching symmetry criterion than conventionalists, the conventionalists agree with the realists about this example of parity reversal.

If T_2 were instead only under $S_3(x,z)=(2x,z)$, the realist would not recognize any overarching symmetry uniting S_1 and S_3, so she would claim that there are important inconsistencies between T_1 and T_2 that warrant further investigation. Such physical symmetries reveal that one of these two theories must be getting something about the world wrong, and one needs a condition on overarching symmetries like consistency that is sensitive to symmetry transformations of empirically identical variables in other physical theories for this interesting result. Conventionalists, however, may argue that there is some overarching symmetry that unites S_1 and S_3, perhaps on the grounds that both physical symmetries transform the position coordinate by a constant integer scaling factor while leaving the other coordinates unchanged, but it is unclear to me what invariance under such an overarching symmetry tells us about the world described by T_1 and T_2. We can no longer say, as we did in the case of parity reversal, that the world doesn't distinguish between left and right since it is not the case that solutions to the

equations of T_2 have parity-reversed counterparts that are also solutions to the equations of T_2. Perhaps the conventionalist takes the interesting metaphysical consequence of invariance under this symmetry to be that there is no metaphysical structure in the world that distinguishes between *either* "left" solutions and "right" solutions *or* solutions and their "position-doubled" counterparts, but such a weird disjunctive structure seems metaphysically worthless. This toy example shows that realist overarching symmetries serve as better guides to both physical and metaphysical projects than do conventionalist overarching symmetries, so one could say that, despite her conventionalism, if a conventionalist has the inclination to draw a useful connection between physical symmetries and ontology, she should be prepared to act as a realist in many situations.

5. Conclusion

Given my approach to physical symmetries, we are left with two ways to proceed: when searching for overarching symmetries, we can follow the realist and search for consistent overarching symmetries, or we can follow the conventionalist and also search for inconsistent overarching symmetries that satisfy some other criterion. As I have argued, conducting the realist's narrow search is more likely to yield the kinds of results that both scientists and philosophers are likely to find pragmatically useful, and as such it may be in the best interest of even the conventionalist to carry out the realist's narrow search first before conducting the broader search.

References

Baker, David. 2010. "Symmetry and the Metaphysics of Physics". *Philosophy Compass* 5 (12):

1157-1166.

Belot, Gordon. 2011 . “Symmetry and equivalence”. In R. Batterman, editor, The Oxford Handbook of Philosophy of Physics, page Forthcoming. Oxford University Press. URL http://philsci-archive.pitt.edu/8446/.

Brading, Katherine and Elena Castellani. 2007. “Symmetry in classical physics. In J. Butterfield and J. Earman, editors, *Handbook of the Philosophy of Physics*, pages 1331– 1367. North-Holland.

Earman, John. 1989. *World Enough and Space-Time: Absolute versus Relational Theories of Space and Time*. MIT Press.

Healey, Richard. 2009. “Perfect symmetries”. *The British Journal for the Philosophy of Science* 60: 697–720.

Ismael, Jenann and Bas van Fraassen. “Symmetry as a guide to superfluous theoretical structure”. In K. Brading and E. Castellani, editors, *Symmetries in Physics: Philosophical Reflections*, pages 371–392. Cambridge University Press, 2003.

Lewis, David. 1983. “New work for a theory of universals”. *Australasian Journal of Philosophy* 61: 343–377.

North, Jill. 2009. “The ‘structure’ of physics: A case study”. *Journal of Philosophy* 106: 57–88.

Olver, Peter. 1993. *Applications of Lie groups to differential equations, 2nd ed*. Springer Verlag.

Roberts, John. 2008. “A puzzle about laws, symmetries, and measurability”. *The British Journal for the Philosophy of Science* 59:143–168.

Defusing Ideological Defenses in Biology

Angela Potochnik

Abstract

Ideological language is widespread in biology. Game theory has been defended as a worldview; sexual selection theory has been criticized for what it posits as basic to biological nature; and evolutionary developmental biology is advocated as an alternative, not addition, to traditional evolutionary biology. Views like these encourage the impression of ideological rift in the field. I advocate an alternative interpretation, whereby many disagreements between camps of biologists reflect unproblematic methodological differences. This interpretation provides a more accurate and more optimistic account of the state of play in the field of biology. It also helps account for the tendency to embrace ideological positions.

1 Ideology and Dissension in Theoretical Biology

Defenders and critics of one or another approach in theoretical biology sometimes employ sweeping, ideologically loaded claims in support of their positions. By this I mean that differences in viewpoint or methodology are construed as resulting from incompatible research programs, each committed to a different view of biological reality. I witnessed one possible result of such a construal a few years ago, when two biologists with different research programs, addressing different types of phenomena, each volunteered an opinion of the other's work. In the view of Biologist A, Biologist B was "no longer doing biology." Biologist

B independently offered the opinion that Biologist A was "not a colleague" of his/hers. Though this was an extreme version of divisiveness, I have witnessed similar exchanges play out in other groups of biologists, both in print and over dinner.[1] Yet these same biologists collaborate in a variety of ways. For instance, Biologists A and B have coauthored publications and shared students. To my mind, this suggests that the presentation of such differences as commitments to fundamentally opposed views of biological reality is ripe for reconsideration. Let us begin by considering three examples of disagreements that have been construed as ideological.

The Optimization Research Program Gould and Lewontin (1979) ushered in an era of polarization in evolutionary biology between "adaptationists" and their critics. In their highly influential paper, Gould and Lewontin explicitly cast as ideology the approach of proposing an adaptive explanation for traits considered individually. They coined an "-ism" for this approach, and they employed religious metaphors to characterize the view. Thus adaptationism "is based on faith in the power of natural selection" and employs the "catechism" that genetic drift is only important in unusual, unimportant circumstances. The adaptationist refuses to credit other causes like drift with any real influence while "[congratulating her/himself] for being such an undogmatic and ecumenical chap." This construal saddles a type of methodology in evolutionary biology with ideological baggage and then criticizes it as false dogma.

Optimization models utilize the procedure that Gould and Lewontin draw into question and are thus one of the primary targets of their criticisms. Many biologists do not accept Gould and Lewontin's ideological gloss of optimization modeling, instead subscribing to Maynard Smith's (1982) interpretation of their point as simply the methodological corrective that optimization models should reflect constraints arising from evolutionary

[1] I do not suggest that such scenarios are more common in biology than other disciplines; the situation in theoretical biology is simply my focus in this paper.

influences besides natural selection. However, a number of defenses of the optimization approach, and evolutionary game theory in particular, have embraced the construal of their position as ideological. Grafen (1984) coined the term "phenotypic gambit" to describe commitment to the optimization approach, which he acknowledges is a "leap of faith." Mitchell and Valone (1990) endorse what they call the "Optimization Research Program," citing Lakatos's view of research programs, the core hypotheses of which adherents should protect from disconfirmation at all costs. Brown (2001) accepts this construal and defends the Optimization Research Program as his "worldview", with game theory at its center. A prominent style of defending optimization modeling thus qualifies as ideological in the sense identified above.

Criticisms of Sexual Selection Theory Sexual selection theory is a well-developed set of hypotheses for the role of selection in the evolution of a variety of sexual and reproductive traits. Different versions of the theory vary in important regards, but I will attempt to give a basic summary that applies to most versions. In many animal species, males (and perhaps sometimes females) are expected to differ in their mating success, which creates selection pressure for traits desirable to members of the opposite sex and/or traits useful in competing with others of the same sex. Thus the peacock's long, colorful train is explained as the result of peahens preferentially mating with comely trained peacocks, not any *survival* advantage conferred by the trains. Similarly, the evolution of combat among male bighorn sheep is explained as the result of ewes preferentially mating with the victors. Traits classically explained as the result of sexual selection range from physical traits, such as ornamentation, to behavioral traits like combat displays or parental care. The basic tenets of sexual selection theory are widely accepted in biology, though as I mentioned there are disagreements about some features, and the hypotheses have been updated and fine-tuned to accommodate ever-expanding information about animals' bodies and behavior (e.g., Clutton-Brock, 2009).

Yet past decades have also seen a number of criticisms of sexual selection theory. Here I will focus on recent criticisms put forth by Roughgarden (2009); see also (Roughgarden, 2004) and (Roughgarden et al., 2006). Roughgarden analyzes and thoroughly rebuts a wide range of hypotheses about the evolution of sex, gender, and reproductive behavior that she attributes to sexual selection theory. Toward the end of the book, Roughgarden argues that she has shown that all those hypotheses are false, that there is no reason to amend the hypotheses, but that sexual selection theory is "a philosophy of biological nature" (p. 246) with an "incorrect foundation." In Roughgarden's view, the hypotheses all "derive from a common view of natural behavior predicated on selfishness, deception, and genetic weeding" (p. 247). Roughgarden suggest that, instead, kindness and cooperation are "basic to biological nature" (p. 1). She thus proposes an alternative "social selection theory," based on the contrary assumptions of "teamwork, honesty, and genetic equality" (p. 247). Roughgarden, then, construes her disagreement with sexual selection theorists as fundamental and expansive, based on beliefs about what is biologically basic. She represents the options as complete commitment to or else complete rejection of all the hypotheses she identifies with sexual selection theory.

Evolutionary Developmental Biology Evolutionary developmental biology, frequently referred to as "evo-devo," is the subfield of biology devoted to studying the evolution of developmental processes. Advocates of evo-devo do not view it simply as an extension of evolutionary biology, but as a needed corrective or even replacement. Müller (2007) contrasts evo-devo with the reigning Modern Synthesis, a synthesis of a number of subfields of biology in the early twentieth century, made possible by the development of population genetics as a way to reconcile discrete Mendelian genetics and gradual evolution by natural selection. According to Müller,

> Whereas in the Modern Synthesis framework the burden of explanation rests on

> the action of selection, with genetic variation representing the necessary boundary condition, the evo-devo framework assigns much of the explanatory weight to the generative properties of development, with natural selection providing the boundary condition. When natural selection is a general boundary condition, the specificity of the phenotypic outcome is determined by development. Thus, evo-devo. . . posits that the causal basis for phenotypic form resides not in population dynamics or, for that matter, in molecular evolution, but instead in the inherent properties of evolving developmental systems (p. 947).

This construes evo-devo not as a supplement to other approaches to evolutionary biology, but as a replacement. The "explanatory weight" goes to development instead of natural selection, for *the* causal basis for phenotypic form is evolving developmental systems, not population dynamics. Carroll et al. (2004) similarly claim that "regulatory evolution is *the* creative force underlying morphological diversity across the evolutionary spectrum" (p. 213, emphasis added). According to Callebaut et al. (2007), evo-devo takes epigenetic considerations as "primordial for the organismic perspective" (p. 41) and thus as providing a "truer picture of life on this earth" (p. 62). As in the two previous examples, advocates of evo-devo present their approach as a view about what is fundamental—in this case, to the evolution of morphology—and the view is a total commitment, in the sense of positing developmental processes as the sole causal basis and hence *the* explanation of these phenomena, to the exclusion of selection.

In each of these debates, the options are presented as sweeping commitments to bipolar positions. Either you subscribe to the Optimization Research Program as your worldview, or you reject it. Either you jettison all of sexual selection theory, or else you commit to the sexual selectionist view of the basics of biological nature. Either you endorse the evolution of developmental systems as the sole causal basis of the evolution of form, or

you unquestioningly uphold the tradition of the Modern Synthesis. These positions are presented as ideological in the sense of involving adherence to a systematic set of ideas, a comprehensive way of looking at things. The set of ideas in question is viewed as fundamental to the domain under investigation, and adherence to one side or the other is taken to be a total commitment. This ideological tenor thus suggests that there is a rift in theory, that there is dispute regarding the basic understanding of these types of phenomena. Here I develop an alternative interpretation, according to which these disagreements and ones like them are more fruitfully seen as rooted in methodological, not ideological, differences (§2). This methodological interpretation provides a more accurate account of how the field of biology functions and a more optimistic take on the state of play in the field. It also suggests a rationale for why some theoretical biologists embrace polarized, ideological positions (§3).

Before proceeding, a couple of clarifications are in order. First, by claiming that these positions are presented as ideological, I do not mean to suggest that they are necessarily influenced by broader *social* ideology. Other research demonstrates that this frequently is the case; Richardson (1984), for instance, develops this point for two of my examples here—game theory and sexual selection theory. Yet the focus of this paper is not the influence of broader social values on theoretical biology, but the construal of debates as ideological in the sense identified above. Second, though I will argue that many debates in biology presented as ideological are more fruitfully understood as methodological debates, this may not hold true for all such debates. Certainly there is room for disagreements in theoretical biology that really do involve commitments to fundamentally opposed positions. One goal of the present analysis is thus to provide resources for distinguishing methodological differences from truly opposed "worldviews."

2 Distinguishing Idealizations from Ideology

There is room for an alternative interpretation of debates in theoretical biology like those surveyed above, despite their ideological tenor. The starting point is philosophical treatments of the role of modeling in science. The scientific practices that have been termed "model-based science" account for the persistence of multiple modeling approaches (e.g. Levins, 1966; Godfrey-Smith, 2006; Weisberg, 2006). On this view, idealized models represent targeted features of a system at the expense of misrepresenting other features. Different modeling approaches thus can *seem* to be incompatible, for they employ different parameters and opposed assumptions, when instead the exact opposite is true. The limitations of idealized models make the use of multiple approaches essential. Taking to heart the idea that models provide a limited representation of only targeted features of a phenomenon makes clear that no single modeling approach offers an exhaustive, fully accurate account of any phenomenon.

This view of model-based science enables an interpretation of seemingly ideological debates in biology as instead methodological at root. Despite the rhetoric sometimes employed, the question to ask about apparently competing modeling approaches is often not which grounds a more successful worldview, but which method better serves one's present research aims. Several aspects of this shift are important. On the methodological interpretation, proposed modeling approaches should be evaluated not according to universal ontological considerations—what the world is posited to be like overall—but considerations of method, especially representational capacity. The evaluation is thus not an absolute judgment, but is contingent on the aims of representation for the research program at hand. This means that different methods may very well be called for in different circumstances, and so a variety of approaches may be warranted. The key features of this interpretation of a debate are thus (1) the resolution depends on evaluation of methodology; (2) choice of approach is contingent on research aims; and (3) multiple approaches can coexist without

	ideological differences	methodological differences
basis of evaluation	what the world is like	method, representational aims
scope of position	complete "worldview"	contingent on research program
commitment to approach	absolute; either/or	multiple approaches can coexist

Table 1: The distinguishing features of ideological and methodological disagreements

difficulty.

These distinguishing characteristics of ideological and methodological disagreements are represented in Table 1. Some disagreements in biology are patently methodological, but many disagreements admit of both construals, including ones traditionally interpreted as ideological. This is so for the three debates I considered above, as I will demonstrate below. There are also some debates for which an ideological construal will remain appropriate. To take an extreme example, embracing basic evolutionary theory commits one to a systematic set of ideas about a type of process and the results it can have. This set of ideas is fundamentally opposed to intelligent design.[2] There is not room for both, for arguments for intelligent design presume the impossibility of evolution. Intelligent design thus cannot be defended on the basis of representational aims.

Let us reexamine the three debates from above, to the end of showing that in each case a methodological interpretation is not only possible but preferable. Although several defenses of the optimization approach have construed the approach as a commitment to a worldview, or a matter of faith, another construal is available. Maynard Smith (1982), for one, attempts to refocus the debate on methodology. This is as strong of a defense as is needed to justify the modeling approaches of optimization and evolutionary game theory, and it is a more defensible position than an ideological defense. Biologists know too much

[2]This example was suggested by a referee for this journal.

about nonselective influences on evolution to subscribe to the notion that selection is the only evolutionary influence. To say that selection is often the only *important* influence, as some have done, is just to declare a preference for tracking that causal process over others. It is more straightforward and more promising to instead defend optimization as simply one modeling approach among many in biology, each with a specific representational focus and delimited range of application.

Mitchell and Valone (1990) represent the debate over the use of game theory as a choice between embracing either the assumptions of evolutionary game theory or those of quantitative genetics, but this is wrong. Certain assumptions of each of these modeling approaches are undeniably idealized, and there are just as obvious limitations to each approach's range of applicability to evolutionary phenomena (Potochnik, 2010). These considerations indicate that game theory and quantitative genetics are each motivated by specific, and limited, representational goals. Each facilitates the faithful representation of some features of some types of evolutionary scenarios. It follows that neither set of assumptions is sufficient for all projects in population biology, which is why both approaches persist. The methodological defense thus better accounts for game theory's role in population biology than does the ideological defense.

The ideological tenor of Roughgarden's (2009) criticisms of sexual selection theory plays an important role for her argument. Advocating the rejection of sexual selection theory in its entirety draws attention to assumptions shared by many of the theory's specific hypotheses, such as competition for mating opportunities and the default traits of each sex, and the regards in which those assumptions may be problematic. Yet a methodological version of Roughgarden's criticisms could still accomplish this. This alternative, methodological approach would be to point out the range of phenomena treated by sexual selection models and assumptions/idealizations the models share. This would set up the desired contrast with Roughgarden's social selection theory, which groups a different range of phenomena

and employs different assumptions. For instance, whereas sexual selection theory addresses scenarios where same-sex animals compete for mating opportunities, social selection theory addresses scenarios where outcomes/selection effects are mediated by social interactions. These groupings of evolutionary phenomena overlap partially, but not entirely. Further, whereas sexual selection theory assumes that direct competition is the norm, social selection theory assumes that a mutually beneficial outcome is within evolutionary reach. It is possible—even likely—that each assumption is right some of the time.

An advantage of this methodological version of Roughgarden's criticisms is that it would provide a less polarizing introduction to the many distinct positive views she advocates, including the alternative modeling techniques she suggests (Potochnik, 2012). Roughgarden lumps her suggestions for modeling approaches together with her complete rejection of sexual selection theory and controversial alternative hypotheses. Faced only with the choice of wholesale rejection or acceptance of those views, many reject them (e.g. Kavenagh, 2006). Yet this need not be so. Roughgarden's suggestions for modeling behavioral evolution, which emphasize malleable selection effects due to influences like negotiation and punishment, are distinct from her specific hypotheses for the evolution of traits related to sex, gender, and reproduction. A methodological approach at once facilitates Roughgarden's criticisms of background assumptions shared by many sexual selection hypotheses and also renders her various ideas separable, and thus potentially palatable to a broader group of biologists.

Evolutionary developmental biology is a valuable field of research, shedding light on an important type of evolution previously neglected by mainstream evolutionary biology. Its focus is how systems of development have evolved, sometimes giving rise to novel features of organisms. To neglect the influence of development on evolved traits and how processes of development have themselves evolved is to ignore an essential element of evolution. This methodological point is sound, and worthy of attention from biologists outside of evo-devo. Yet the idea voiced by advocates of evo-devo that developmental systems are the *sole* causal

basis for phenotypic form, and that natural selection is merely a "boundary condition" (Müller, 2007), is going too far in the opposite direction. Evolution is an incredibly complex, prolonged process, with a variety of important causal influences that combine and interact in myriad ways. Different modeling approaches will capture different elements of that process and employ simplifying assumptions and idealizations to exclude other elements. They will also apply more aptly to different ranges of evolutionary phenomena. Evo-devo draws attention to one set of causal influences, viz., developmental processes, that are especially important for certain types of evolution, viz., morphological evolution. This provides an important part of the evolutionary story, but it does not *replace* the stories that instead feature natural selection (or drift, etc.) Shifting from an ideological to a methodological defense thus would be a valuable change for advocates of an evo-devo approach as well. As with the earlier two examples, evo-devo can be motivated more effectively when practitioners of other methods are not asked to declare a new worldview.

These examples of disagreements about biology thus can be profitably interpreted as rooted in methodological differences, despite the tendency of many biologists to construe the differences as ideological in nature. The same is true for other debates in biology that are similarly structured, such as the longstanding disagreements surrounding group selection. Recall that I do not expect *all* apparently ideological debates to be resolved on methodological grounds. Instead, each debate must be examined to see whether it can be construed to possess the features of a methodological disagreement, as summarized in Table 1. On the methodological interpretation, competing approaches should not be evaluated according to which is true, or the basis of a successful worldview, and a complete commitment to an approach is unwarranted. The evaluation is instead based on which types of systems and which features of those systems are central to one's present research program, and which approach best meets those representational aims.

It is important to note that, even when a methodological interpretation is appropriate,

there still may be disagreements about matters of fact. For instance, two biologists may well disagree on whether natural selection is a significant causal factor in the evolution of a particular trait. But such disagreements need not amount to universal commitments, and they are not the only reason for variation among biologists' methods. The methodological interpretation of disagreements in theoretical biology keeps models' aims and limitations at center stage, which results in the evaluation of an approach contingent on the aims of research and the likelihood of the coexistence of multiple approaches in a stable area of research.

3 Normal Science with a Twist

Features of this methodological interpretation of debates can actually help account for why some biologists on each side of these issues embrace polarized, ideological positions. In the section above, I suggested that research programs within biology differ in ways that warrant employing certain modeling approaches to the exclusion of others. For central as well as accidental reasons, participants in different research programs focus on different phenomena; are acquainted with different bodies of past research; and even may have familiarity with different varieties of organisms. This means that advocates and critics of a modeling approach address that approach from different locations, for they often differ in both interests and expertise. Such differences can easily lead to disagreements about the commonness of types of phenomena and the significance of causal patterns. Those engaged in optimization research are well familiar with the successes of optimal foraging theory, and they dismiss the overdominance of malaria-resistance as an uncommon if not unique genetic situation. Roughgarden's hypotheses lead her to focus on animal species with extensive social interactions, such as shared care of young or collective hunting. And evo-devo theorists are well familiar with the evolution of limbs.

Another ingredient of ideological stances in theoretical biology is an implicit commitment to the existence of simple causal processes with broad domains of application. A tacit belief in such "magic bullet" causes enables differences in focus and expertise among researchers to be interpreted as commitments to different types of causes. If it is agreed that most phenomena are influenced by a vast array of causal factors, then researchers' differences are naturally understood to arise from a difference in focus, not a difference in worldview. In this case, the claim that certain features of the evolutionary process are more important is reduced to the claim that some are worthier of investigation than others. Put this way, it is not an empirical claim, but merely a statement of research interests (see Godfrey-Smith (2001) on this point regarding adaptationism in particular).

This account of how ideological positions in theoretical biology arise in a sense explains away such ideological tendencies. Yet I should emphasize that the posited account attributes more significance to ideological positions than, say, the idea that these stances are simply adopted as a way to increase recognition or funding. In my view, standoffs between opposed ideological positions indicate something important about the field of biology. That there are such entrenched proponents and opponents of different methods indicates that a variety of modeling approaches have some purchase on the evolutionary process and other biological phenomena. In my view, this reflects the complex causal processes at work in biology, and the endless variety in how causal factors combine and interact. There are evolved traits like foraging behavior that optimization analysis readily predicts; those like sickled red blood cells with which it can get nowhere; and a whole range of intermediate traits for which it is partially successful insofar as it represents the causal contribution of natural selection, which may be just one causal influence among many. The causal influences on social behavior in animals are likely as diverse as the behaviors themselves, so there is room for sexual selection theory's success with some behaviors and failure with others. Development and evolution are both without question causal influences on organisms' traits; how these influences interact

is just as certain to be highly variable.

Recasting ideological differences as methodological differences also grounds a more optimistic interpretation of the current state of play in theoretical biology. The diversity of approaches does not stem from a clash of worldviews, and so biology is not in a state of crisis from which one research program will emerge triumphant. Instead, strong ideological differences persist within a functional field of research. This will continue to be the case so long as different methodologies are useful in different research programs.

So, then, why does the main point of this paper matter? If ideological differences are consistent with a fully functional field of science, why concern oneself with the reinterpretation I suggest? In my view, were more biologists and philosophers of biology to embrace this interpretation of commitment to favored modeling approaches, real, advantageous consequences would result. Most basically, less attention would be devoted to unnecessary arguments that are, as it turns out, about preferred phenomena and modeling approaches of choice. A prime example is the decades of continuing debate in philosophy of biology over adaptationism, when optimization approaches can instead be motivated on much more modest grounds (Potochnik, 2009).

Adopting the methodological interpretation would also promote cooperation among those who continue to have substantive disagreements about biology. Instead of becoming mired in ideological impasse, focusing on modeling approaches allows communication and progress in spite of different views about how the models apply to the real world. Godfrey-Smith claims that,

> When much day-to-day discussion is about model systems, disagreement about the nature of a target system is less able to impede communication. The model acts as a buffer, enabling communication and cooperative work across scientists who have different commitments about the target system (2006, p. 739).

On this view, even continuing disagreements about evolutionary phenomena need not hinder

cooperative work on features of models. If all parties can, at least temporarily, set aside differences in commitment to broad claims of causal importance, they can further joint understanding of models' inner workings and conditions of their application. Indeed, I have observed this first-hand at meetings of a working group on evolutionary game theory (at the National Institute for Mathematical and Biological Synthesis).

Finally, the refocus facilitated by a shift to the methodological interpretation of disputes in biology creates more room for activities of significance for theoretical biology and the philosophical analysis of biology. Recognition of the viability of a range of modeling approaches and the related idea of complex and variable causal processes should lead to a diminished focus on isolated, illustrative applications of a type of model. This should be replaced by an increased focus on determining the range of and conditions for a modeling approach's applicability and the limitations of its assumptions, as well as increased attention to the interplay among multiple causal influences. For philosophers of biology, the lesson is to expect a continual plurality of methods in biology—methods that can appear contradictory—and to take with a grain of salt any claim that one or another approach is the key to understanding biology.

Acknowledgments

Thanks to Rob Skipper for helpful advice on this project and to Francis Cartieri for his valuable feedback and research assistance. I am also grateful to three anonymous referees for *BioScience* for their insightful comments. This work was completed with the help of a summer research fellowship from the University Research Council at the University of Cincinnati.

References

Brown, Joel S. (2001), "Fit of form and function, diversity of life, and procession of life as an evolutionary game", in Steven Hecht Orzack and Elliott Sober, eds., *Adaptationism and Optimality*, Cambridge Studies in Philosophy and Biology, Cambridge: Cambridge University Press, chap. 4, 114–160.

Callebaut, Werner, Gerd B. Müller, and Stuart A. Newman (2007), "The Organismic Systems Approach: Evo-devo and the Streamlining of the Naturalistic Agenda", in Roger Sansom and Robert N. Brandon, eds., *Integrating Evolution and Development: From Theory*, Bradford, chap. 2, 25–92.

Carroll, Sean B., Jennifer Grenier, and Scott Weatherbee (2004), *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*, Wiley, 2nd ed.

Clutton-Brock, Tim (2009), "Sexual selection in females", *Animal Behaviour* 77: 3–11.

Godfrey-Smith, Peter (2001), "Three Kinds of Adaptationism", in Steven Hecht Orzack and Elliott Sober, eds., *Adaptationism and Optimality*, Cambridge Studies in Philosophy and Biology, Cambridge: Cambridge University Press, chap. 11, 335–357.

——— (2006), "The strategy of model-based science", *Biology and Philosophy* 21: 725–740.

Gould, Stephen Jay, and R.C. Lewontin (1979), "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme", *Proceedings of the Royal Society of London, Series B* 205: 581–598.

Grafen, Alan (1984), "Natural selection, kin selection, and group selection", in J.R. Krebs and N.B. Davies, eds., *Behavioural ecology: An evolutionary approach*, Oxford: Blackwell Scientific Publications, chap. 3, 2nd ed.

Kavenagh, Etta (Ed.) (2006), "Letters: Debating sexual selection and mating strategies", *Science* 312: 689–694.

Levins, Richard (1966), "The strategy of model building in population biology", *American Scientist* 54: 421–431.

Maynard Smith, John (1982), *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.

Mitchell, William A., and Thomas J. Valone (1990), "The Optimization Research Program: Studying Adaptations by their Function", *The Quarterly Review of Biology* 65: 43–52.

Müller, Gerd B. (2007), "Evo-Devo: Extending the Evolutionary Synthesis", *Nature Reviews: Genetics* 8: 943–949.

Potochnik, Angela (2009), "Optimality modeling in a suboptimal world", *Biology and Philosophy* 24: 183–197.

——— (2010), "Explanatory Independence and Epistemic Interdependence: A Case Study of the Optimality Approach", *The British Journal for the Philosophy of Science* 61: 213–233.

——— (2012), "Modeling Social and Evolutionary Games", *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 202–208.

Richardson, Robert C. (1984), "Biology and Ideology: The Interpenetration of Science and Values", *Philosophy of Science* 51: 396–420.

Roughgarden, Joan (2004), *Evolution's Rainbow: Diversity, Gender, and Sexuality in Nature and People*, Berkeley: University of California Press.

——— (2009), *The Genial Gene: Deconstructing Darwinian Selfishness*, Berkeley: University of California Press.

Roughgarden, Joan, Meeko Oishi, and Erol Akçay (2006), "Reproductive social behavior: Cooperative games to replace sexual selection", *Science* 311: 965–969.

Weisberg, Michael (2006), "Forty years of 'The Strategy': Levins on model building and idealization", *Biology and Philosophy* 21: 623–645.

Human Nature in a Post-Essentialist World

Grant Ramsey
Department of Philosophy
University of Notre Dame

1. Introduction. In a Platonic worldview, which species an individual belongs to is answered by this question: in which form does it participate? A goat is a goat because it participates in the Platonic form goat and not, say, the form horse. The tidiness of this answer, however, belies the questionable ontology on which it rests. It is for this reason that contemporary philosophers generally eschew the invocation of Platonic forms in producing a theory of species natures.

Nevertheless, species seem to have a nature—perhaps not one founded on Platonic forms, but instead on a set of essential properties. If we were given a lineup of a random assortment of vertebrates—wildebeest, aardvarks, naked mole rats, vampire bats, and humans—we would have no problem picking out the humans. Even if the lineup were populated with our closest living relatives, we would have no difficulty picking the humans out from the other apes. This is true in part because humans differ in many ways from other species of ape. But it is not merely that humans differ from the other related species, it is also that individual humans share many traits amongst themselves. This similarity—this set of traits that it seems we are able register intuitively to instantly recognize an individual as a human being—is, it would seem, what we could use as a respectable foundation for human nature: human nature is just that set of traits that are possessed by to each individual and essential to their being human.

But with Darwin's (1859) publication of *On the Origin of Species*, an essentialist view of species was called into question. Darwin argued that the history of life has a tree

structure and that branches on this tree represent biological taxa. A large branch represents a high-level taxon, such as a class or phylum, whereas a small branch represents a small taxon like a genus or species. For Darwin, species are not ontologically sui generis. Instead, there is a continuum from variations within species to genuine specieshood. On this view, taxonomists debate the number of species within a taxon—flowering plants, for example—not (merely) because there is a dearth of taxonomic data, but because there are (often well-justified) differing approaches to drawing a line in the continuum between variations and species.

Darwin's view of species and the origin of species through intraspecific variation is mirrored by contemporary biology. The question at hand, then, is whether the concept of species nature—and in particular human nature—has any place in this contemporary theoretical context. In what follows, I will examine two ways in which philosophers have answered this question. I will then show that neither answer is satisfactory and will then present my own alternative.

2. Hull's skepticism about human nature. The main line contemporary taxonomic framework, cladistics, recognizes and formalizes Darwin's insights about the tree structure of life. For cladists, the only legitimate taxa are monophyletic clades, which are groups formed by encircling an ancestor and all of the branches descended from this ancestor. Taxonomic groupings that include multiple clades (polyphyletic groups) or fail to include all of the branches within a clade (paraphyletic groups) are not legitimate taxa. Thus, a species must be monophyletic, and not polyphyletic or paraphyletic. A corollary

of these restrictions is that in order for a new taxon to arise, there must be a branching event. No new taxa, no new species, genera, etc., can arise in the absence of branching.

It is this aspect of contemporary taxonomy that Hull (1986) uses to make his argument that there is no such thing as human nature. Hull's arguments can be summed up in the following way: 'Human nature' must pick out intrinsic traits that are exhibited by all (and only) humans. This set of traits must be definitive of and essential to membership in the species *Homo sapiens*, just as having eight protons is definitive of and essential to being oxygen. But membership in *Homo sapiens*, as with membership in any biological species is determined not by essential properties shared by each individual, but instead by their position within a clade. Thus, such essential properties cannot be definitive of species membership. Additionally, these properties are unlikely to be exhibited by all *Homo Sapiens* (individuals with severe developmental disorders are members of our species, after all). Further, even if a synchronic time slice of *Homo sapiens* reveals an interesting set of shared traits, the species will keep evolving and these traits are unlikely to persist over the entire existence of the species.

One way to understand Hull's argument is that he takes a species to be an individual whose birth and death are marked by phylogenetic branching events (nodes). The organisms belonging to the species are thus all of the organisms existing between the nodes. Hull's synchronic argument says that a synchronic slice in this species is unlikely to produce a set of individuals with interesting traits shared by them and only them. And the diachronic argument goes further to say that even if the synchronic slice produces interesting traits, these traits are probably not going to persist through all intra-nodal time.

3. Machery's reply and its limitations. In the face of Hull's skepticism, one could either concede that there really is no such thing as human nature, or produce a concept of human nature that eschews or challenges Hull's criticisms. Machery (2008) attempts the latter. He argues that while Hull's arguments are devastating to essentialist notions of human nature, they do not undermine another concept of human nature, what he is labeling the "nomological" view. Machery's nomological view holds that "human nature is the set of properties that humans tend to possess as a result of the evolution of their species" (323).

Machery's account addresses Hull's worries in part by no longer considering human nature to be definitional. Thus, because an individual is not defined as being a human, that is, belonging to the species *Homo sapiens*, in virtue of possessing the traits that fall under the rubric of human nature, particular individuals can lack one or more of these traits while still being human. Instead, Machery merely requires that any trait considered part of human nature must be possessed by most humans.

Despite the successful dodge of Hull's arguments, there are difficulties with Machery's view. First, by requiring that the trait be possessed by the majority of humans, one looses many traits characteristic of humans. Any traits (psychological, behavioral, morphological) that are sexually dimorphic or, say, exhibited only by a particular ethnic group, will be excluded. Vivipary, lactation, and menopause, for example, are no part of human nature. Machery recognizes this, but holds to his view because "saying that humans have a nature entails that humans form a class that is of importance for biology.

The members of this class tend to have some properties in common in virtue of evolutionary processes" (326).

But why does belonging to almost all humans make it an important class for biology? Is it not a biologically interesting feature of human nature that the females undergo menopause? Furthermore, why should we presume that it is the *sameness* across individuals that is of interest to scientists, and not their *variation*? As Geertz (1973) insightfully pointed out, "[t]he notion that unless a cultural phenomenon is empirically universal it cannot reflect anything about the nature of man is about as logical as the notion that because sickle-cell anemia is, fortunately, not universal, it cannot tell us anything about human genetic processes. It is not whether phenomena are empirically common that is critical in science […] but whether they can be made to reveal the enduring natural processes that underly them" (44). I am in full agreement with Geertz on this point—it a mistake to hold that only traits universal (or nearly universal) in the human species are of scientific interest and should be included within human nature.

The second difficulty is that Machery takes it to be unproblematic to sort properties into two bins, those due to "the evolution of their species" (323) and those "exclusively due to enculturation or to social learning" (326). Only the former, asserts Machery, are a part of human nature. But what, exactly, is a property "exclusively due to enculturation or to social learning"? Any organismic property is going to be due both to heritable features of the organism as well as the particular environmental features the organism happens to encounter during its life. Some of these environmental features could be counted as instances of "enculturation" or "social learning," but the fact that such environmental features are present in the organism's life history does not mean that

we can point to properties as being "exclusively due" to these environmental inputs. The innate-acquired dichotomy has been long challenged (see Lehrman 1953; Bateson and Mameli 2007) and I see no way to make Machery's distinction without a futile attempt at reifying this problematic dichotomy.

Thus, although Machery successfully dodges Hull's criticisms, the concept of human nature that he ends up proposing accords neither with intuitive notions of human nature, nor with scientific practice. It should, therefore, be discarded and be replaced. In what follows, I will propose and defend a replacement. The aim of my alternative account of human nature will be to fulfill core desiderata for such a concept. Human nature should (D1) be the empirically accessible (and thus not based on occult essences) subject of the human (psychological, anthropological, economic, biological, etc.) sciences, (D2) help clarify related concepts like *innateness* and *naturalness*, which are associated with human nature, and (D3) characterize human uniqueness. Although for some, an additional desideratum that human nature tell us something about what humans *should be* or *should strive for* is important, the notion of human nature that I offer will not directly fulfill this normative desideratum, since I hold that human nature cannot simultaneously fulfill this desideratum and D1.

4. The life history trait cluster account of human nature. Consider an individual human of a particular genetic constitution in a particular environment. There are many different possible outcomes to such an individual's life. Think of these possible outcomes as possible life histories. One life history involves the individual becoming relatively prosperous, having a large family, and dying after a long life. Another life history

terminates from a fatal disease in childhood. These life histories are populated by a multitude of traits. Some of these traits persist over entire life histories (e.g., a core body temperature in excess of 90°F), while others are short lived (e.g., a temperature of 105°F), or momentary (a particular sneeze). 'Trait' here thus picks out any feature of a life history, no matter its duration or significance.

Now consider the totality of traits and how they are dispersed over the totality of possible life histories. It is clear that the traits do not populate the possible life histories in a random way—instead, there are patterns: Some later ("consequent") traits will always or usually follow certain prior ("antecedent" traits). The consequent trait of speaking fluent English will always be linked with antecedent exposure to spoken English. This might seem trivially true, but it is important that this is not the case with all individuals (human or otherwise). The traits can of course be morphological or physiological or psychological and not just behavioral. For individuals lacking the gene for the enzyme phenylalanine hydroxylase (PAH), those with (and only with) antecedent exposure to phenylalanine will exhibit the consequent trait of phenylketonuria (PKU), a devastating neurological disease. This rather simple observation that traits are non-randomly dispersed over this set of life histories (and that there will be robust patterns of antecedent-consequent pairs) provides the basis for the notion of individual nature, from which I will construct the concept of human nature. *Individual nature* is defined as the pattern of trait clusters within the individual's set of possible life histories. This concept of individual nature, though coherent, points to something that scientists will typically want to study only as a means to learning about humans in general—learning, that is, about human nature.

Human nature is defined as the pattern of trait clusters within the totality of extant human possible life histories. Thus, if one were to take all of the possible life histories that form the basis for individual nature, and then combine them, one would possess the set of life histories that forms the basis for human nature. The trait distribution patterns in this set of life histories constitute human nature. For example, the traits "bearing offspring" and "lactating" will be clustered, the former being antecedent to the latter. There will be few instances of lactating that are not associated with the antecedent trait "bearing offspring." It is this pattern of human life history trait clusters that I am identifying with human nature.

Note that I am not arguing that these patterns are what *define* membership in the species. I am thus not adopting a homeostatic property cluster conception of the definition of species (see Boyd 1999). The trait patterns in an individual's nature do not determine to which species the individual belongs (though they of course serve as evidence); it is instead, as Hull and the scientific consensus argues, belonging to a particular lineage that determines species membership. This view of human nature is thus fully consistent with the cladistic view of specieshood.

This account of human nature I will label the Life-history Trait Cluster (LTC) account. This is to distinguish it from accounts that are essentialist or normative, since it is neither based on essential properties, nor, as we will see in section 7, does it imply that human nature is in any sense "good." Instead, characterizations of features of human nature are merely descriptions of statistical trends within the collective set of human life histories.

5. Human nature: the subject of the social and psychological sciences. At first blush, the LTC account of human nature might seem to be of little use. It fails to identify a property or set of properties essential to (or good for) being human. Furthermore, by linking human nature to an infinite set of life histories, it would appear that human nature is not even empirically accessible, thus failing D1. In this section, I hope to show that, on the contrary, the LTC account identifies just what it is that is the subject of the human sciences.

To begin, let us consider what kind of results of psychological studies are of value. A study merely reporting that humans are sometimes aggressive will be of little interest. In the LTC framework, it is uninformative because it is merely calling attention to the existence of some traits within the set of human life histories, but is not identifying (or quantifying) a pattern of these traits. Such a study would never be published in an academic journal. If, instead, the study reports that adults who were abused as children will tend to be aggressive toward their own children, then the study is of interest and, if executed well, is the sort of research that could be published. What such a study is doing is identifying a pattern in the collective life histories. It is making the claim that life histories with the antecedent trait "abused as a child" will tend to be associated with the consequent trait "aggressive toward one's children."

Similarly, controlled experiments are seeking to discover life history trait patterns. A study that has participants give speeches on unfamiliar topics in front of an unfriendly audience and then measures cortisol levels in their saliva are searching for such patterns. Here a possible pattern would be an antecedent "uncomfortable public speaking event" followed by "high cortisol levels." Control groups in such studies are used to see whether

the presence of the antecedent is causally linked to the consequent. And in the LTC framework, the controls are a way of refining our knowledge of the trait patterns and projecting into unknown possibilities. The psychologist wants to state, quite generally, that public speaking is a source of stress. And making these general statements is to say that there is a robust pattern of association between the antecedent and consequent traits. Thus, knowledge of human nature (in the LTC sense) is just the aim of psychological investigations such as these.

By extension, it is easy to see that knowledge in the human sciences more generally is, for the most part, knowledge of human nature. An anthropologist who describes an unusual behavior among the Yaminawa is going to investigate both the meaning of the behavior as well as its causes. Such an investigation is but an investigation into trait patterns—what other psychological, behavioral, physiological, etc. traits are linked to the unusual behavior? Similarly, the behavioral economist who shows that greater choice leads to poorer satisfaction is pointing to a life-history trait pattern: antecedent states "decisions with many options" will be associated with consequent states like "poor satisfaction." The degree to which the findings are robust is the degree to which there is a strong statistical association between the antecedent and consequent in the set of life histories.

Non-human animals are often used as models for studying humans. The reason why such research can be useful is also accounted for by the LTC. A model (a rat, say) in some domain (like cancer research) is going to useful to the extent that the same antecedent-consequent pattern exists in both humans and rats—cancer as consequent and extensive exposure to benzene as antecedent, for example.

6. The quantification of human nature. We have seen that D1 is satisfied: human nature is indeed the subject of the human sciences. But it has not yet been made clear what the LTC account implies about other ways that the concept of human nature is used, i.e., it has not yet been shown to fulfill D2 (the clarification of related concepts like innateness and naturalness). We speak of a behavior being "natural," or it being part of "human nature" to behave in a particular way. Are such locutions undermined by the LTC account or can they be understood within it? In this section I will show that while it is a mistake to understand traits as dichotomously either "natural" or not, a part of "human nature" or not, I will show that there is a sense in which traits can be more or less natural, more or less central to human nature. In order to accomplish this, I will construct a human nature space and suggest that behaviors occupying a particular region are core features of human nature, while those in other parts of the space are less central.

Human nature, as argued for above, is investigated by determining associations between antecedent and consequent traits in the collective human life histories. There are two key variables that one could use in characterizing these associations. First, there is the proportion of life histories that exhibit the antecedent trait. This could also be understood as the probability that an individual drawn at random will exhibit the antecedent during their life. Second, there is the proportion of those exhibiting the antecedent who also exhibits the consequent. This can also be understood probabilistically as the conditional probability of exhibiting the consequent given the antecedent. I will label the first the pervasiveness, p, of the antecedent, and I will label the second the robustness, r, of the antecedent-consequent association. Some antecedent

traits (lacking PAH and consuming phenylalanine) will be rare, but robustly associated with their consequent (PKU). And some antecedents can be common (imbibing alcohol) but not very robustly associated with particular consequents (like esophageal cancer), despite the fact that imbibing alcohol does raise the incidence of such cancer.

These examples exhibit two important features of the *p*-*r* space. One is that the antecedent need not be a single, simple trait, but can instead be complex a trait or cluster of traits. The second is that there will generally be a tradeoff between *p* and *r*. For a given consequent, one can often increase robustness by adding more antecedent traits (or replacing a simpler antecedent with a more complex one). Lacking the gene for PAH will be associated with PKU, but the realization that individuals can consume diets absent in phenylalanine makes the absence of PAH not all that robustly linked with PKU. However, the antecedent "lacking PAH + consuming phenylalanine" is more robustly linked to the consequent, PKU. The same is true of the alcohol example. Singling out heavy drinkers, or heavy drinkers that are also smokers, will increase the robustness of the link between the antecedent and esophageal cancer.

This tradeoff between *p* and *r* parallels the tradeoff that Lewis (1973) saw in the creation of deductive systems. He argued that "a contingent generalization is a law of nature if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength" (73). Such deductive systems can generally be axiomatized more or less simply, and there is a tradeoff: "Simplicity without strength can be had from pure logic, strength without simplicity from (the deductive closure of) an almanac" (73). By making the antecedent traits more and more complex, one can increase *r*, but at a cost of decreased *p*. And one can achieve a

high *p* by simplifying the antecedent, but this usually comes at a cost to *r*. The analogy of Lewis's almanac is a set of antecedents that picks out the totality of antecedent traits. The consequent would have perfect robustness, but the specific set of antecedents will be singular, with the lowest possible pervasiveness.

With the *p*-*r* space in mind, we are now in the position of being able to return to the question of what it might mean to "behave naturally" or for a behavior to be part of human nature. The LTC framework implies that instead of saying that it is natural to C, we should instead say that it is natural for As to C, where 'A' denotes the antecedent(s) and 'C' denotes the consequent. Thus, instead of saying that it is natural to develop PKU, one should instead say that it is natural for those who (A) lack PAH and consume phenylalanine to (C) develop PKU. Similarly, instead of saying that PKU is a part of human nature, it is more informative and precise to say that it is part of human nature for individuals who lack PAH and consume phenylalanine to develop PKU. Analogously instead of stating that "lactation is a part of human nature," one could state that "lactation is a part of female human nature" or that "female lactation is a part of human nature," since the latter descriptions pick out an antecedent (being female) that makes the antecedent-consequent association robust.

If a trait is a part of human nature, then so, too, it might seem that it is innate. As discussed above, the innate-acquired distinction is problematic. But there is nevertheless a sense in which the LTC framework can provide a revised and improved way of understanding this concept. Here are three ways that one could understand innateness in the LTC framework. First, the *r*-value of the trait association is one possible way of quantifying innateness—the higher the *r*-value, the more the consequent is associated

with the given the antecedent. Thus, it is not that consequent traits, or traits in general, that are innate full stop, but they have a quantifiable degree of innateness provided the antecedent. Second, one could restrict innateness to antecedent-consequent associations that exhibit both a high *r*-value and a high *p*-value.

Third, although these understandings of innateness harmonize with some of the standard ways of understanding the concept (in terms of canalization, for example), they do not preserve the "innateness = not learned" definition, since learning can be a part of the causes of the consequent. To preserve the "not learned" conception of innateness, one could add the restriction that learning must not be causally relevant to the appearance of the consequent, given the antecedent, though I imagine that many traits one would be apt to call innate would no longer be classified as such under this restricted definition, since learning is woven into the causal fabric of so much of development.

The LTC account of human nature thus fulfills D2 and, as we saw in the previous section, D1. But what about D3, namely, the identification of human uniqueness? And, furthermore, is there any sense in which the framework can provide insight into human goodness—what we should strive for in becoming a good human, or what we should aim for through human enhancement?

7. Human uniqueness and the question of normativity. The LTC account is incredibly permissive. All sorts of antecedent-consequent links exist, many of which are rather trivial. It is human nature for females to lactate, but this is true of all mammals. This does not mean that it is not an important feature of human nature, but it does mean that it is not uniquely human. And there are countless rather trivial trait associations. "Every human

that has mass will die" has maximal *r* and *p*-values, but is utterly otiose. In fact, trait associations with maximal *r* and *p*-values will tend to be trivial. The interesting ones often occur when the antecedent is not universal, and when changes in the antecedent are causally associated with changes in the consequent. In order to both eliminate the trivial associations and to capture human uniqueness, I will define *uniquely human nature* as the subset of the antecedent-consequent associations that are unique to the human species. Importantly, uniquely human nature, like human nature, is a property not of each individual human, but instead of the set of extant human (actual and possible) life histories. Speaking a human language fluently (provided exposure to this language while growing up) is part of uniquely human nature. This is true because no other species will speak fluently given a similar upbringing. Raising a chimp in the same way as an American child does not result in it speaking English. The same will be true of many of the consequents that we laude in the human species, such as complex systems of morality and the ability for self-reflection.

The LTC account's "uniquely human nature" is thus a way of capturing human uniqueness, satisfying D3. But what of normativity, is there a sense in which human nature is good, or can be improved upon via human enhancement? The short answer is that because the foundations of the LTC framework are trait distribution patterns, it is, strictly speaking, descriptive and not normative. Furthermore, there is not some eternal "human nature," like a fixed target in Plato's heaven, that humans can strive for. Instead, human nature simply tracks the form, behavior, etc. of humans. Human nature was different in our species' past and will be different in the future.

But this does not mean that there are no moral implications of human nature under the LTC framework. If the study of human nature is the study of patterns of trait associations, then studying human nature will provide insight into human goodness and evil—if a particular nefarious consequent is robustly associated with a particular set of antecedents, then the elimination or reduction of the consequent should be pursued via the elimination or alteration of one or more of the antecedents. Similarly good consequent traits can be made more common via an increase in the antecedents with which they are robustly associated.

8. Conclusions. "*Man, in a word, has no nature*" (Ortega y Gassett, 1961, 217, emphasis in original). Such a sentiment is shared by many in the humanities and human sciences, and seems to be based on the reflection that humans are simply too diverse (across cultures, genders, times) for there to be some human essence that we could extract from this diversity. Such skepticism is warranted if the only notion of human nature on the table is an essentialist one. But the LTC framework provides an alternative. It embraces the diversity, showing that there are patterns within and across human heterogeneity. If there is to be an empirically-accessible human nature that sheds troubling essentialisms, then it should be founded on the unique pattern of traits within the collective human life histories. Such a concept of human nature cannot play all of the roles that we may desire of it—showing us how to be more fully human, for example—but it can play many of the other roles. I have shown that it can be understood as the subject of the various human sciences, can clarify what we mean when we classify a trait as innate or natural, and can also provide a basis for human uniqueness.

References

Bateson, P. and Mameli, M. 2007. The innate and the acquired: Useful clusters or a residual distinction from folk biology? *Developmental Psychobiol*ogy, 49: 818–831

Boyd, R. 1999. Homeostasis, species, and higher taxa. In *Species: New interdisciplinary essays.*, ed. Robert A. Wilson, 141-185MIT Press. Cambridge.

Darwin, C. 1859. *On the origin of species by means of natural selection* London: Murray.

Geertz, C. 1973. *The interpretation of cultures* Basic Books.

Hull, D. L. 1986. On human nature. Paper presented at PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 2: 3-13.

Lehrman, D. S. 1953. A critique of konrad lorenz's theory of instinctive behavior. *The Quarterly Review of Biology* 28 (4): 337-63.

Lewis, D. K. 1973. *Counterfactuals* Harvard University Press.

Machery, E. 2008. A plea for human nature. *Philosophical Psychology* 21 (3): 321-9.

y Gasset, J. O. 1961. *History as a system: And other essays toward a philosophy of history* WW Norton & Company.

Biol Philos
DOI 10.1007/s10539-012-9344-0

Viral information

Forest Rohwer · Katie Barott

Received: 16 April 2012 / Accepted: 24 September 2012

Abstract Viruses are major drivers of global biogeochemistry and the etiological agents of many diseases. They are also the winners in the game of life: there are more viruses on the planet than cellular organisms and they encode most of the genetic diversity on the planet. In fact, it is reasonable to view life as a viral incubator. Nevertheless, most ecological and evolutionary theories were developed, and continue to be developed, without considering the virosphere. This means these theories need to be to reinterpreted in light of viral knowledge or we need to develop new theory from the viral point-of-view. Here we briefly introduce our viral planet and then address a major outstanding question in biology: why is most of life viral? A key insight is that during an infection cycle the original virus is completely broken down and only the associated information is passed on to the next generation. This is different for cellular organisms, which must pass on some physical part of themselves from generation to generation. Based on this premise, it is proposed that the thermodynamic consequences of physical information (e.g., Landauer's principle) are observed in natural viral populations. This link between physical and genetic information is then used to develop the Viral Information Hypothesis, which states that genetic information replicates itself to the detriment of system energy efficiency (i.e., is viral in nature). Finally, we show how viral information can be tested, and illustrate how this novel view can explain existing ecological and evolutionary theories from more fundamental principles.

Keywords Virus · Phage · Information · Ecology · Evolution

F. Rohwer (✉) · K. Barott
Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA
e-mail: frohwer@gmail.com

Published online: 31 October 2012

Viruses as information

Viruses are the only biological entities that replicate purely as information. When a virus enters its host, the virion completely disassembles and the nucleic acid is copied into new genomes, which are then packaged and released as new virions. Physically, there is nothing in the original form of the virion that has to be passed on from one generation to another. Not one single molecule, atom, or quark must be transferred between the old and new. The only thing that must be moved between viral generations is the information to build the next set of viruses. The rest of biology operates differently. Every new cell physically shares all of its molecules with the original mother cell at the moment of division.

Here we propose the Viral Information Hypothesis, which argues that:

(1) Physical information is about position in the Universe.
(2) Biology creates physical information by changing the position of matter, effectively working as Maxwell's Demon.
(3) Viral information converts different types of physical information into itself at the cost of overall energetic efficiency.
(4) There is a thermodynamic cost to destroying physical information, which is quantified by Landauer's Principle. Extremely large populations like viruses experience selection at the Landauer limit and this is observable.

Welcome to the viral world

Humans observe nature at the meso-scale (e.g., mm to km). Our brains are good at processing this sort of data, from observing blossoming cherry trees, to scuba diving on a coral reef, to measuring the beaks of finches. But our senses have led us astray because until recently we have been overlooking most of life. On the cherry tree's blossom, roots, branches, and leaves are millions of viruses and their microbial prey. When swimming over a coral reef, every milliliter of seawater is home to ten million viruses (Bergh et al. 1989; Hara et al. 1991; Fuhrman 1999) and every surface, including the mucus on corals and fish, is covered by even more viruses (Wilson et al. 2005; Marhaver et al. 2008; Patten et al. 2008; Willner et al. 2010). And much of the DNA flying about in Darwin's famous finches actually belongs to microbes and viruses.[1]

Viruses are particularly easy to overlook because they are completely outside our sensory range. This is a problem, because by missing the virosphere biologists have effectively ignored the most abundant and diverse biological entities on Earth. Conservatively, there are 1.0×10^{31} of them. This is based on estimates of $\sim 1.0 \times 10^{30}$ microbes on the planet (Whitman et al. 1998) and an average of ~ 10 viruses per prokaryotic cell (Weinbauer 2004). An alien visiting our planet, given a different sensory range that could directly detect viruses, would likely consider

[1] Doing some back-of-the-envelope calculations for the zebra finch, *Taeniopygia guttata*, we estimated that about 10 % of the total DNA in a finch is microbial or viral.

them the dominant form of life. (Note to reader: if you are fluent in the history and biology of viruses, feel free to skip the following section as we review these topics.)

How do we know that there are this many viruses? Initially, they were counted using electron microscopes (Bergh et al. 1989). Now they are routinely counted using epifluorescent microscopy (Noble and Fuhrman 1998). For example, to enumerate the viruses in a milliliter of seawater, the sample is pulled through a glass filter with 0.02 micron pores (small enough to capture viruses). Then the filter is treated with a DNA stain that lights up under fluorescent light on the microscope. Technically, what biologists actually count are virus-like particles (VLPs). A VLP is something that looks like a virus but has not formally been characterized and shown to act like a virus; that is to infect and then replicate inside a host cell. Even though viruses are incredibly small, 10^{31} make them a huge crowd. If you line up all the viruses in single-file, the line would reach a thousand times across our home galaxy.

While the total number of viruses is enormous, what is really incredible is their dynamics (Weinbauer 2004). Our best estimates are that every week 10^{31} viruses fall apart and 10^{31} new ones are made to replace them. This means that roughly 1.7×10^{25} new viruses are produced every second. For each new virus, approximately 50,000 base pairs of DNA have to be synthesized (Steward et al. 2000). Thus, each second more than 10^{30} base pairs of viral DNA are made on planet Earth. Since the vast majority of these viruses infect microbes (Bacteria and Archaea, two of the three domains of life), the making of these viruses entails the death of approximately 10^{24} microbial cells each second. This enhances the microbial diversity and productivity of ecosystems. It also is a huge factor in global energy and nutrient cycling (Fuhrman 1999). The point of these exercises is to show just how numerous, massive, and dynamic these 10^{31} viruses really are. When considering the virosphere, extremely unlikely events become probabilistic certainties.

Even though viruses dominate our home in the universe, most people consider them only when they cause some sort of disease.[2] But in fact, most viruses are actually phage: viruses that infect the Bacteria. In 1915 the Englishman Frederick Twort discovered an "ultra-microscopic virus" that converted bacteria into fine granules (Twort 1915). In his usage, the word 'virus' seems to have meant simply an infectious agent. He wrote about "a minute bacterium that will only grow on living material…or a form of life more lowly organized than the bacterium" (1915: 1242). The virus was destroyed at 60 °C and could not be cultured except on the bacteria. "On the whole it seems probable, though by no means certain, that the active transparent material is produced by the micrococcus, and since it leads to its own destruction and can be transmitted to fresh healthy cultures, it might almost be considered as an acute infectious disease of micrococci" (1915:1243). That is, the bacteria were getting sick.

The French-Canadian Felix d'Herelle went further and showed that a filterable "antagonistic microbe" capable of killing the bacteria *Shigella dysenteriae* was

[2] Attempts to identify the bacteriological pathogen of diseased tobacco plants led to the identification of the first virus by Beijerinck in 1898. The infectious agent was described as a filterable *contagium vivum fluidum,* and was later named the tobacco mosaic virus (Bos 1999).

isolatable from patients who developed enteritis following dysentery infections (d'Herelle 1917). He performed the first plaque assays and showed that titers of this agent were highest during patient recovery. Culturing the agent required living dysentery bacteria, but under these conditions the agent could be cultured through 50 successive transfers. d'Herelle wrote: "the disappearance of the dysentery bacilli is coincident with the appearance of an invisible microbe…This microbe, really a microbe of immunity, is an obligate **bacteriophage**"—the first use of the term (d'Herelle 1917: 159).

Bacteriophage means "bacteria eating" and is usually shortened to *phage*. They are a subclass of viruses that infect the Bacterial domain of life (Woese et al. 1990). The early virus hunters quickly realized the phage were very diverse, with each one finicky about which host it would infect. This specificity of phages for specific strains of bacteria was one of the early ways of microbiological identification, which was carried out through a procedure called phage typing (Williams and Rippon 1952). Basically, this can be done by culturing bacteria in a test tube and then adding different phage. If the tube clears, it means all the bacteria have been killed by the phage. Using this approach, thousands of phage and their host strains have been classified. Determining host range is one of the most a useful approaches for characterizing the virosphere.

Another way to characterize viruses is to visualize them with an electron microscope (EM) (Bradley 1965). Viruses outside a host cell are called virions and they are some of the most wondrous creatures ever discovered. The archetypical phage looks like a lunar lander, with a protein capsid protecting the double stranded DNA genome (Fig. 1). It has a tail, which is used to transfer the viral genome into the host cell, and tail fibers that help the phage find the correct host (Fig. 1). Viral capsids usually form one of two basic architectures: rods or icosahedra, the size of which can vary by more than an order of magnitude, and which can package genomes that also differ more than 30-fold in size (Fauquet et al. 2005) (Fig. 1). In forming rod-shaped particles, the capsid proteins are arrayed in a helix around the viral DNA or RNA. TMV (tobacco mosaic virus) is the classic example of this shape (Klug 1999). The other common capsid structure is an icosahedron surrounding a nucleic acid core (Fauquet et al. 2005). Among the five thousand-plus phages that had been described and viewed under the EM by 2000, 96 % are "tailed phages" (Ackermann 2001), composed of an icosahedral head containing the genome, and possessing a tail that functions to identify the host and deliver the genome to the cell interior. The tail structures divided the lunar-lander phage into three main groups: the lambda-like phage, which have long, flexible tails; the T7-like phage, which have short, contractile tails; and the T4-like phage, which have long contractile tails (Ackermann 2007). Often the capsid proteins and genomic nucleic acids self-assemble in vitro to form infectious virions (Hung et al. 1969; Kushner 1969; Lebeurier et al. 1977; Klug 1999). Neither additional information nor an energy source is required. Viruses that infect animals and plant often have the icosahedral structure enclosed in an envelope of lipids (Fig. 1).

The virions are exquisitely designed predators that seek out and kill their hosts. Overall, the virions have a slightly negative charge so that they repel each other when the host cell is lysed (Todd 1927; Krueger et al. 1929; Clifton and Madison

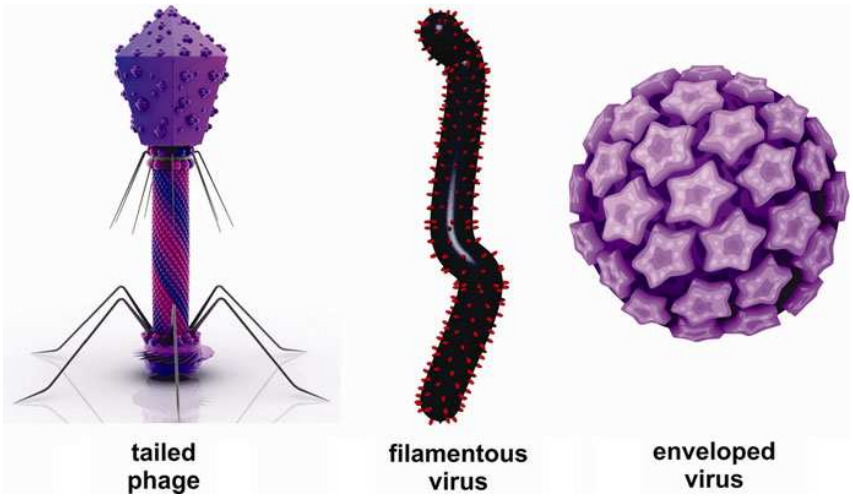


Fig. 1 Examples of the main types of viruses: tailed phage that infect bacteria, filamentous viruses that infect all domains of life, and enveloped viruses that infect animal and plant cells. There are actually hundreds of variants on these basic themes and interested readers should look at the International Committee on Viral Taxonomy (ICTV) website and/or Viral Taxonomy books. Of particular interest are the numerous novel virions associated with Archaea viruses

1931). This allows them to spread out and avoid entanglement in the lysed cell's released contents. More subtly, it appears that the charges are arranged so that the virions are actually dipoles (De Groot et al. 1977); that is they have a negative charge around the capsid head and a slightly positive charge at the tails (Serwer and Pichler 1978; Kosturko et al. 1979). This presumably orientates them tail-first when making an attack on the bacterial host cell (which is slightly negative).

During the attack phase, the virion is first electrostatically attracted to the cell's surface (Krueger 1931; Delbrück 1940). It rolls along the outside and the tail searches for specific receptors. If the host is the correct bacterial species, then the phage will find the receptor and clamp down on it (Heller 1992). When this happens the phage's tail will drill through host cell's membranes and cell wall so that the viral genome can be delivered into the cytoplasm (Letellier et al. 2004). To achieve this, the outside of the contractile tails rearrange their molecular structure so that the tube inside the sheath can pierce the cell (Kanamaru et al. 2002; Leiman et al. 2004). This allows for the DNA to be injected with incredible force (Kindt et al. 2001; Letellier et al. 2004). The process is not dissimilar to the secondary jaws of Ridley Scott's Aliens; terrifying if you happen to be the size of a microbe.

In addition to tails, the phage capsids are often decorated with secondary structures that facilitate the attack. This includes hooks that grab hold of bacterial flagella so that the phage is pulled down to the host (Schade et al. 1967; Lotz et al. 1977). Other molecular accessories probably help the virion survive different environmental conditions or act as camouflage to throw off protective ectoenzymes produced by the host. The constant war between the virion's capabilities for finding and infecting the cell, and the retaliation by the host, leads to evolutionary dynamics

known as Red Queen (Van Valen 1974) and ecological cycles called Lotka-Volterra/Kill-the-Winner (Bratbak et al. 1990).

There were several problems that had to be circumvented in order to study the diversity and dynamics of the global virome. To culture a virus you need to grow its host and at the present time we only routinely cultivate roughly 1 % of the microbes from the environment (Fuhrman and Campbell 1998). And once conditions to culture the microbe are identified, they have to be modified to encourage infection by a virus. Because of these challenges, the culturing route would be a daunting and defeating path to take. What about sequencing the viral DNA? Sequencing of the 16S ribosomal RNA gene (rDNA) is a common technique used to analyze the diversity of microbial communities, and it capitalizes on the high conservation of this one gene amongst all microbes, thereby avoiding comparison of entire genomes to get at community diversity (Pace et al. 1986; Woese 1987). However, it not possible to take a similar approach with viruses because there is no gene in common between all groups (Rohwer and Edwards 2002). To get around this limitation, a technique for shotgun sequencing random fragments from the pool of all of the viral genomes in the community was developed (Breitbart et al. 2002). This approach is called metagenomics.

Analysis of the entire genetic pool (the metagenome) of a sample was first performed on viral communities isolated from seawater in San Diego (Breitbart et al. 2002). This early study showed that the vast majority of viral sequences (80 % or more) were not recognizable using common bioinformatic searches. That is to say that the uncultured viral DNAs were so dissimilar from every single known sequence accumulated in various databases of known viral, bacterial, and eukaryotic sequences (e.g., GenBank) that we have no idea what they do or to whom they belong. Despite the incredible volume of sequences added to the databases since these metagenomes were first sequenced, most viruses remain unknown. On the other hand, microbial metagenomes, which followed closely behind the first viral metagenomes, were much less mysterious with <20 % of sequences not matching anything in the databases (Dinsdale et al. 2008). Because viruses are incredibly abundant, much more so than microbes, and because the majority of the information contained in viral genomes is unknown, viruses are the final frontier of unexplored genomic diversity and are the largest genetic repository that exists. We are left with the question: *why are there so many viruses*?

Demons and information

Up until this point, we have argued that viruses are extremely abundant, incredibly diverse, and travel through time and space as information. We propose that this relationship between viruses and information is the key to their success, but what does "information" mean? In the communication sense, information is a measure of "surprisal" (Tribus 1961). The greater the surprise at observing an object, the more information it contains. Gold contains more information than does hydrogen (i.e., it is more surprising to find gold). Considering this concept more deeply, it becomes clear that information is actually an accounting of position in the universe. That is,

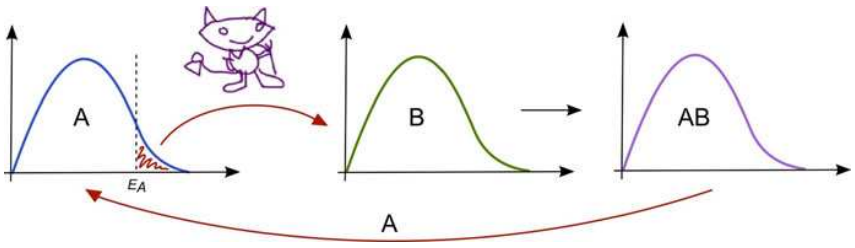


Fig. 2 Illustration of Maxwell's Demon and Landauer's principle. The Demon/enzyme selectively picks "A" molecules with sufficient energy to react with reactant "B", which leads to product "AB". This process slightly cools the "A" population. This loss of heat is put back into the system by the surrounding Universe. During degradation/erasure of "AB", "A" goes back into its population and this heat can be measured using methods like isothermal calorimetry

the gold is created by compressing protons, neutrons, and electrons together in space and time. As these particles become locked together, degrees of freedom are lost and a highly unlikely, and therefore a highly informative event, is created. *This organizing of matter is time and space is physical information.*

Physical information does not come for free. The thermodynamic consequence of physical information was first mathematically defined by Rolf Landauer, who calculated that the minimum energy (E) stored in one bit of information was equal to $kT\ln(2)$, where $k =$ Boltzmann's constant and $T =$ temperature in Kelvin (Landauer 1996). Heat released by the erasure of physical information can best be envisioned by invoking Maxwell's Demon. Originally presented as a challenge to the Second Law of Thermodynamics, the Demon is a hypothetical creature that can pick the "hot" molecules from one container and mover them to another. This creates a temperature differential, which could be used to drive some sort of engine. So, with the right Demon, we can create a perpetual motion machine. It was Leo Szilard who showed that the reason this does not happen is because the Demon is actually gaining information about the relative position of the molecules (Szilard 1929). This realization killed the perpetual motion machine and Maxwell's challenge to the Second Law of Thermodynamics.

Now consider a Maxwell's Demon in a biochemical system (Fig. 2). At a certain temperature, the reactant molecules "A" have different velocities, as described by Boltzmann's distribution. The fastest/hottest "A" molecules are on the right side of the distribution. For our purposes, the molecules above the activation energy (E_A) are the ones with sufficient velocity to be active in a chemical reaction. Now imagine a Maxwell's Demon that selectively picks "A" molecules within the E_A population and passes them to a second reactant pool "B". This creates the product and effectively traps both molecules in product "AB". In doing so, the demon has increased the information of the system. When "AB" degrades into its components, "A" will re-enter the original population and heat it up.[3] This increase in

[3] Sometimes it is easier to think of this as only two molecules of A. When one molecule is taken away, the system gets colder. And when the A molecule returns to the system, it gets hotter.

temperature is described by Landauer's Principle (Landauer 1996; Toyabe et al. 2010).

We propose that biology behaves as Maxwell's Demon, where the Demons are enzymes that selectively grab E_A molecules to form products. This creates physical information, which can be used to do work (Toyabe et al. 2010; Bérut et al. 2012). The one caveat to this *work-from-information* schema is that it requires elaborate scaffolds like a computer. We suggest that genetic information is the set of instructions to construct the scaffold for Maxwell's Demons such that they convert different types of physical information into more instances of itself. This new information has a thermodynamic cost when it is erased and the amount of heat released by the destruction of information is also described by Landauer's Principle (Landauer 1996; Toyabe et al. 2010). It should be possible to observe the link between physical information and thermodynamics and use it to better understand biology and in particular the success of viruses.

Viral versus physical information

Let us compare and contrast physical and viral information. Gravity organizes the physical properties of the universe. Gravity clumps matter, which enhances the importance of the other three fundamental interactions. By organizing matter in time and space, gravity creates physical information. The cloud of subatomic particles from the Big Bang could have spread out evenly throughout the universe. Instead, small imperfections allowed gravity to pull some particles together; and these attracted others. Accretion discs developed and collapsed into stars, where gravity fused the matter together forming heavier elements and led to the production of electromagnetic radiation. These processes increase the physical information content within the universe through strictly physical processes (Fig. 3a). Gravity also reinforces itself—bigger things attract more objects, creating a positive feedback loop. Biology also reinforces itself by organizing compounds and concentrating them. Just like gravity, life creates organization of particles in the universe through juxtaposition and rearrangement. The organization of matter by biology leads to viral information because it converts physical information into itself at the cost of maximal efficiency from a thermodynamic point of view.

Viral information and the rest of biology

The two primary sources of physical information used for conversion into viral information are electromagnetic radiation from fusion (the basis for phototrophy) and the redox byproducts of fission (the basis for chemotrophy) (Fig. 3a). Let us examine the largest biome on earth (potentially), the deep, hot biosphere that exists within the Earth's crust (Gold 1992; Chapelle et al. 2002), as an example of how the viral information feedback might work. In this ecosystem, a very simple energy source, split water, provides the energy to create cellular biomass. At the temperature and pressure of this system, the only known predators of the microbial

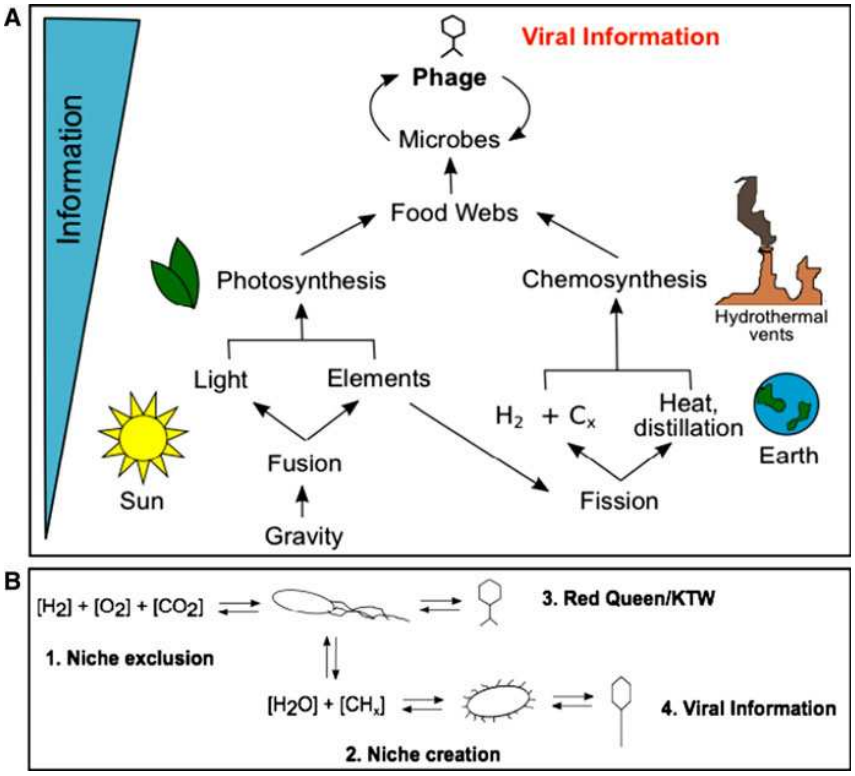


Fig. 3 From gravity to viral information: dust to phage. **a** Schematic of how gravity leads to viral information. **b** Schematic of how viruses shape ecology (1-3) and evolution (3), leading to diversification and an increase of viral information (4)

inhabitants are the viruses. From this simple food web, the main "rules" of ecology and evolution are apparent (Fig. 3b). These are (1) niche exclusion, (2) niche creation, and (3) Red Queen/Kill-the-Winner (KTW) dynamics, which ultimately result in and are driven by (4) viral information.

In the first step, one microbial population makes a living by using up resources from the local environment (e.g., split water). This leaves the system depleted of these items, generating competition and niche exclusion; the microbe that exploits these resources the fastest wins. At the same time, the viruses in the system essentially punish the most successful microbe by killing it (Bratbak et al. 1990; Rodriguez-Brito et al. 2010). Viral lysis releases cellular debris into the surrounding environment, and the new microbes that capitalize on this new set of resources then begin the process anew, excluding others from their new niche, creating a new set of waste products and resources, and feeding a new population of viruses. Effectively, the viruses are creating conditions to replicate themselves.

The pressure of predation also leads the microbe to alter parts of itself to avoid viral recognition (i.e., Red Queen Dynamics, or running to stay in the same place),

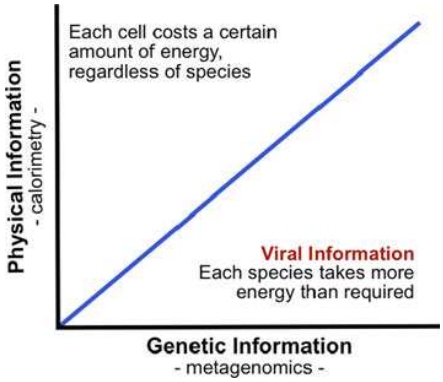


Fig. 4 Searching for viral information. The *line* indicates where the amount of physical information and genetic information contained within cells are equal. Communities above the line contain more physical than genetic information due to low genetic diversity (few species but many individuals), with each individual requiring a certain amount of energy regardless of its genetic composition. Communities below the line contain more genetic information. It is here that the energetic cost of information becomes apparent, and where we expect to find viral information

while the virus adapts to recognize the new microbes. In other words, viruses drive the evolution of microbial genomes and niches, ultimately leading to the increase of viral information. Since viruses can sample more sequence space, they wind up generating the greatest amount of new genetic information, which can then be passed along to the host microbes through horizontal gene transfer (HGT). Ultimately, the community is converting physical information into genetic information. We hypothesize that this step is viral because it is done at a great thermodynamic inefficiency; that is a lot of waste heat is produced. Using the rule of thumb that each trophic transfer loses 90 % of the heat, each joule of viral information gained costs the system 100 J of physical information.

Measuring viral information

The destruction of physical information, as discussed above, results in the release of heat according to Landauer's Principle. This heat can be measured by calorimetry. Specifically, isothermal calorimetry tells us about the conversion rate of physical information of a community into heat.[4] Genetic information of the same community can also be measured, in this case by sequencing the DNA. Based on these two techniques, we propose the following experiment where physical information is followed using calorimetry, and genetic information is followed using metagenomics. When the two are plotted as shown in Fig. 4, we propose that a community dominated by viral information occurs in the lower right region of the graph where

[4] In this sense, physical information is the 1/S, where S is entropy. We prefer to use "information" for our accounting of position, because it better explains the concept. However it is completely compatible to express this in terms of S.

genetic information is made at the expense of thermodynamic efficiency (i.e., low conversion to physical information).

Is there any evidence that viral information is real? Djamali and colleagues used isothermal calorimetry to study the heat released by marine microbial and viral communities (Djamali et al. 2012). In this experiment, viruses lowered the standing stock of the cellular component by ~25 %. At the same time, viruses increased the work output of the system by over 200 %. The decline in cell numbers coupled with the increase in diversity looks very much like viral information. Future experiments of this type offer a framework for testing the Viral Information Hypothesis.

Observations of viral information in nature

As one possible example of the consequences of viral information in nature, let us consider the global conservation of viral sequences. Specific PCR "hunts" for the same virus and/or virally encoded genes have shown that those viruses/viral genes are relatively common all over the world (Breitbart et al. 2004; Short and Suttle 2005; Casas et al. 2006). For example, PCR primers were designed to specifically to amplify two viral sequences named HECTOR and PARIS (Breitbart et al. 2004). These so-called PUP sequences (Polymerases from Uncultured Podophage) were present in most environments investigated and were found to be essentially identical (>99 % conserved at the nucleotide level). Similarly, metagenomic samples have found exactly identical, overlapping viral sequences from widely dispersed parts of the ocean (Angly et al. 2006). Finally, genomic sequencing of phage has identified exactly matching sequences in very different phage genomes (Graham Hatfull, personal communication).

The widespread occurrence of nearly identical sequences across the planet requires an explanation. We hypothesize that this extremely faithful global conservation is due to the energy cost associated with information erasure. As we have seen, there are literally an astronomical number of viruses on the planet. It is estimated that each viral population (that is the number of individuals of the same species) is 10^{23}. If each virus in a population has a difference of one bit of information, then the heat released by destroying that additional information would be 1,800 J via Landauer's Principle (Fig. 5). In other words, a viral population that has one mutation per genome replication costs 1.8 kJ more to replicate than a viral population that has no mutations. Over the course of a year, the amount of energy required by the mutating viral population versus the non-mutating population is approximately 100 kJ, assuming that the whole population is replaced once a week. Over a billion years, this is 10^{14} J, which is roughly equivalent to the amount of energy released by an atomic bomb.[5] Energetically efficient populations outcompete those that are less efficient (Meysman and Bruers 2007; Vallino 2010); therefore, populations of viruses with reduced mutations rates will outcompete those with higher mutation rates, all other things being equal.

[5] Remember this competition is actually occurring locally for small parts of the total population.

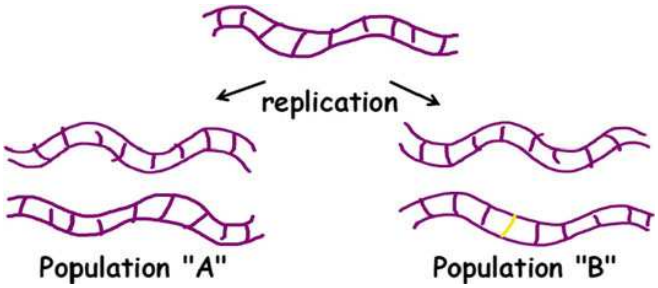


Fig. 5 The Landauer limit and mutations. A mutation in a DNA population creates at least 2 bits of Physical Information. It costs an extra 3–6 $\times$ 10^{-21} J to erase the "B" population

The extra energetic costs of physical information associated with a mutation might explain why identical viral sequences are observed on a global scale. Physical information in the sense of a mutation is an extremely small selection pressure and *we hypothesize that the Landauer limit is the smallest force of selection*. Because of their "information only" life styles, it is easier to observe the thermodynamic consequences of information in viral communities.[6] Furthermore, we only observe this in Nature because it is extremely hard to raise 10^{23} phage (or any other biological entity) in the laboratory.

Conclusion

Envisioning the biosphere as a massive system that ultimately feeds viruses helps us address a major outstanding question: why is biological diversity dominated by viruses? This question would not have even occurred to earlier biologists, simply because they did not know the extent of the virosphere. Modern biology, however, needs to incorporate this natural phenomenon into its canon. The Viral Information Hypothesis has the potential to synthesize ecology and evolutionary theory by incorporating the viruses with the rest of biology in a thermodynamic framework.

Acknowledgments The authors would like to thank their Long-Suffering Listeners—Anca Segall, Peter Salamon, Jim Nulton, Ben Felts, and Beltran Rodriguez-Brito—who have provided a ready and critical audience for these ideas. Merry Youle edited some early versions of this manuscript, and anonymous referees provided many helpful comments. Thank you Willow Segall for some of the art work.

[6] The viruses may not be the only place where Landauer's limit can be observed. Ultraconserved Elements (UCE) occur in plant and animal genomes. There are 481 UCEs (200 bp) between human, mouse and rats, which represents 300 million years of evolution. Reneker et al. identified large numbers of UCEs (>100 bp) shared between plants and animals (Reneker et al. 2012).

References

Ackermann HW (2001) Frequency of morphological phage descriptions in the year 2000. Arch Virol 146:843–857

Ackermann H-W (2007) 5500 phages examined in the electron microscope. Arch Virol 152:227–243

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F (2006) The marine viromes of four oceanic regions. PLoS Biol 4:e368

Bergh Ø, BØrsheim KY, Bratbak G, Heldal M (1989) High abundance of viruses found in aquatic environments. Nature 340:467–468

Bérut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R, Lutz E (2012) Experimental verification of Landauer/'s principle linking information and thermodynamics. Nature 483:187–189

Bos L (1999) Beijerinck's work on tobacco mosaic virus: historical context and legacy. Phil Trans R Soc Lond B 354:675–685

Bradley DE (1965) The morphology and physiology of bacteriophages as revealed by the electron microscope. J R Microsc Soc 84:257–316

Bratbak G, Heldal M, Norland S, Thingstad TF (1990) Viruses as partners in spring bloom microbial trophodynamics. Appl Environ Microbiol 56:1400–1405

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 99:14250–14255

Breitbart M, Miyake JH, Rohwer F (2004) Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol Lett 236:249–256

Casas V, Miyake J, Balsley H, Roark J, Telles S, Leeds S, Zurita I, Breitbart M, Bartlett D, Azam F, Rohwer F (2006) Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California. FEMS Microbiol Lett 261:141–149

Chapelle FH, O'Neill K, Bradley PM, Methe BA, Ciufo SA, Knobel LL, Lovley DR (2002) A hydrogen-based subsurface microbial community dominated by methanogens. Nature 415:312–315

Clifton CE, Madison RR (1931) Studies on the electrical charge of bacteriophage. J Bacteriol 22:255–260

d'Herelle FH (1917) Sur un microbe invisible antagoniste des bacilles dysenteriques. Comptes rendus Acad Sci 165:373–375

De Groot G, Greve J, Block J (1977) Transient electric birefringence of the bacteriophages T3 and T7. Biopolymers 16:639–654

Delbrück M (1940) Adsorption of bacteriophage under various physiological conditions of the host. J Gen Physiol 23:631–642

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632

Djamali E, Nulton JD, Turner PJ, Rohwer F, Salamon P (2012) Heat output by marine microbial and viral communities. J Non-Equilib Thermodyn 37:291–313

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA et al (2005) Virus taxonomy: VIIIth report of the international committee on taxonomy of viruses. Elsevier Academic Press

Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399:541–548

Fuhrman JA, Campbell L (1998) Microbial microdiversity. Nature 393:410–411

Gold T (1992) The deep, hot biosphere. Proc Natl Acad Sci USA 89:6045–6049

Hara S, Terauchi K, Koike I (1991) Abundance of viruses in marine waters: assessment by epifluorescence and transmission electron microscopy. Appl Environ Microbiol 57:2731–2734

Heller KJ (1992) Molecular interaction between bacteriophage and the gram-negative cell envelope. Archives Microbiol 158:235–248

Hung PP, Ling CM, Overby LR (1969) Self-assembly of Qbeta and MS2 phage particles: possible function of initiation complexes. Science 166:1638–1640

Kanamaru S, Leiman PG, Kostyuchenko VA, Chipman PR, Mesyanzhinov VV, Arisaka F, Rossmann MG et al (2002) Structure of the cell-puncturing device of bacteriophage T 4. Nature 415:553–557

Kindt J, Tzlil S, Ben-Shaul A, Gelbart WM (2001) DNA packaging and ejection forces in bacteriophage. Proc Natl Acad Sci USA 98:13671

Klug A (1999) The tobacco mosaic virus particle: structure and assembly. Phil Trans R Soc Lond B 354:531–535
Kosturko LD, Hogan M, Dattagupta N (1979) Structure of DNA within three isometric bacteriophages. Cell 16:515–522
Krueger AP (1931) The sorption of bacteriophage by living and dead susceptible bacteria. J Gen Physiol 14:493–516
Krueger AP, Ritter RC, Smith SP (1929) The electrical charge of bacteriophage. J Exp Med 50:739–746
Kushner DJ (1969) Self-assembly of biological structures. Bacteriol Rev 33:302–345
Landauer R (1996) The physical nature of information. Phys Lett A 217:188–193
Lebeurier G, Nicolaieff A, Richards KE (1977) Inside-out model for self-assembly of tobacco mosaic virus. Proc Natl Acad Sci USA 74:149–153
Leiman PG, Chipman PR, Kostyuchenko VA, Mesyanzhinov VV, Rossmann MG (2004) Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host. Cell 118:419–429
Letellier L, Boulanger P, Plançon L, Jacquot P, Santamaria M et al (2004) Main features on tailed phage, host recognition and DNA uptake. Front Biosci 9:1228–1339
Lotz W, Acker G, Schmitt R (1977) Bacteriophage 7–7-1 adsorbs to the complex flagella of Rhizobium lupini H13–3. J Gen Virol 34:9–17
Marhaver KL, Edwards RA, Rohwer F (2008) Viral communities associated with healthy and bleaching corals. Environ Microbiol 10:2277–2286
Meysman FJR, Bruers S (2007) A thermodynamic perspective on food webs: quantifying entropy production within detrital-based ecosystems. J Theor Biol 249:124–139
Noble RT, Fuhrman JA (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquat Microb Ecol 14:113–118
Pace N, Stahl D, Lane D, Olsen G (1986) The analysis of natural microbial populations by rRNA sequences. Adv Microb Ecol 9:1–55
Patten NL, Harrison PL, Mitchell JG (2008) Prevalence of virus-like particles within a staghorn scleractinian coral (Acropora muricata) from the Great Barrier Reef. Coral Reefs 27:569–580
Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu C-R, Korkin D (2012) Long identical multispecies elements in plant and animal genomes. Proc Natl Acad Sci USA. doi:10.1073/pnas.1121356109
Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-Cuadrado AB, Mira A, Nulton J, Pal[scaron]li[cacute]l L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber RV, Willner D, Youle M, Rohwer F (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4:739–751
Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol 184:4529–4535
Schade SZ, Adler J, Ris H (1967) How Bacteriophage χ attacks motile bacteria. J Virol 1:599–609
Serwer P, Pichler ME (1978) Electrophoresis of bacteriophage T7 and T7 capsids in agarose gels. J Virol 28:917–928
Short CM, Suttle CA (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl Environ Microbiol 71:480–486
Steward GF, Montiel JL, Azam F (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnol Oceanogr 45:1697–1706
Szilard L (1929) über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. Zeitschrift für Physik A Hadrons and Nuclei 53:840–856
Todd C (1927) On the electrical behaviour of the bacteriophage. Br J Exp Pathol 8:369
Toyabe S, Sagawa T, Ueda M, Muneyuki E, Sano M (2010) Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. Nat Phys 6:988–992
Tribus M (1961) Thermostactics and thermodynamics. D. van Nostrand, Princeton
Twort FW (1915) An investigation on the nature of ultra-microscopic viruses. The Lancet 186:1241–1243
Vallino JJ (2010) Ecosystem biogeochemistry considered as a distributed metabolic network ordered by maximum entropy production. Phil Trans R Soc B 365:1417–1427
Van Valen L (1974) Molecular evolution as predicted by natural selection. J Molec Evol 3:89–101
Weinbauer MG (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev 28:127–181
Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 95:6578–6583

Williams REO, Rippon JE (1952) Bacteriophage typing of Staphylococcus aureus. J Hyg (Lond) 50:320–353

Willner D, Desnues C, Rohwer F (2010) Viral metagenomics: from fish slime to the world. In: Li RW (ed)Metagenomics and its applications in agriculture, biomedicine, and environmental studies. Nova Scientific Publishers, Hauppauge, p 337–366

Wilson WH, Dale AL, Davy JE, Davy SK (2005) An enemy within? Observations of virus-like particles in reef corals. Coral Reefs 24:145–148

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Was Leibniz the First Spacetime Structuralist?

Abstract

I argue that the standard interpretation of Leibniz as a relationist about space is mistaken, and defend a reading according to which his correspondence with Samuel Clarke actually suggests that Leibniz holds a view closely resembling modern spacetime structuralism. I distinguish my proposal from Belot's recent reading of Leibniz as a *modal* relationist, arguing for the superiority of my reading based on the Clarke correspondence and on Leibniz's conception of God's relation to the created world. I note a tension between my proposal and Leibniz's ontology, and suggest that a solution is forthcoming and worth pursuing.

1. Introduction. The canonical reading of Leibniz's view of space and time holds that he was a thoroughgoing *relationist*: roughly, he believed that there is nothing to space over and above the various relations of coexistence between bodies, and he believed that there is nothing to time over and above the relations of succession between events. This reading dates back to Russell and is perhaps recapitulated most fully in Earman's *World Enough and Spacetime* (1989); recently, Gordon Belot has suggested a more nuanced variant of it. Importantly, the received view relies very heavily on a correspondence between Leibniz and Samuel Clarke, in which Leibniz seems to argue transparently and at great length for the relationist conception of space that has long been attributed to him. I believe that this reading reveals a misunderstanding of what Leibniz says about space in the *Correspondence*. My goal in this paper, accordingly, will be to reconstruct in a somewhat schematic way what Leibniz's remarks therein actually tell us about his theory of space.

In a nutshell, I believe that his actual view looks suspiciously like a modern view known as *spacetime structuralism*, and my investigation will revolve around the claim that a plausible reconstruction of his view of space indicates that he was, for all intents and purposes, a proto- spacetime structuralist. In other words, Leibniz held a view of space very similar to that held by the modern spacetime structuralist, though he formulated it in different terms and based it upon his own particular metaphysics. I will proceed in the following manner, then: I'll first situate the canonical reading of Leibniz in light of a quick reconstruction of the main tenets of Newtonian substantivalism. Next, I'll introduce and explain spacetime structuralism, providing background for my discussion of Leibniz's views.

After this, I'll launch the promised investigation of Leibniz's view of space as he presents it against Clarke. In the course of the investigation, I'll distinguish my reading of Leibniz from Belot's and motivate a rejection of Belot's reading in favor of mine. At the very least, I hope to show how the machinery of spacetime structuralism enhances our understanding of Leibniz's view. But what I really want to establish is that Leibniz was, in a sense, the first spacetime structuralist: the lineage of this hotly debated view goes back much further than one would have thought.

2. Newtonian Substantivalism and Spacetime Structuralism. Let's now examine Newton's view of space. In the first Scholium of the *Principia*, Newton provides perhaps his most concise statement of what has come to be known as "substantivalism", saying that "absolute space, of its own nature and without reference to anything external, always remains homogeneous and immovable", and that "place is that part of space that a body occupies" (2004, 64-65). Space, in other words, exists over and above bodies; it's a preexisting "container" that would still be there even if there were no bodies. It is, in Earman's words, "a substratum of points underlying physical events" (1989, 10). Space and its parts "maintain their own identities independently of physical bodies", to quote a recent paper by John Roberts (2003, 555). The essence of the Newtonian view is that the parts of space – i.e. points – possess intrinsic identity. Now, the standard reading of Leibniz on space commits him to the outright denial of Newton's claim: space does *not* exist prior to, or over and above, physical bodies *in any sense*; the parts of space not only lack intrinsic identity but aren't even

properly thought of as locations within a substantival container. This is the essence of what has come to be known as “relationism”. Earman puts the claim this way: “spatiotemporal relations among bodies... are direct; that is, they are not parasitic on relations among a substratum of space points that underlie bodies” (1989, 12). On the standard reading, Leibniz's positive claim about space emerges from the negative claim in that space is simply the order of bodies, and nothing more, and would not exist without bodies.

Now, as I've said, I'm proposing that Leibniz's *actual* conception of space becomes clear when viewed through the lens of spacetime structuralism, and that there's a good deal of evidence that he was actually a proto- spacetime structuralist himself. As background for this interpretation, we need to recall the views of the spacetime structuralist. Broadly speaking, spacetime structuralism is an instance of a more general view in the philosophy of science called *ontic structural realism*, which is roughly the idea that, in Esfeld's and Lam's words, “there are objects, but instead of being characterized by intrinsic properties, all there is to [them] are the relations in which they stand” (2008, 31). The view amounts to the claim that the relational complexes described by fundamental physics fully individuate the relata that they contain; these relata include things like electrons and spacetime points. Wuthrich summarizes the view (without advocating it) in a recent paper: “The objects... do not have any intrinsic properties but *only relational ones*. So what is really there... is a network of relations among objects that do not possess any intrinsic properties but are purely defined by their 'place' in [a relational structure]” (2009, 1042). One can also distinguish, as Wuthrich does, two broad variants of the view: one according to which objects and relations are

ontologically on a par with each other, and another according to which what's fundamentally real is just the set of relations, and objects are only thought of as somehow emerging from those relations. This distinction will become important in my discussion of Leibniz's view.

The spacetime structuralist applies some version of ontic structural realism to the case of general relativity. The individuals in the domain of general relativity – the individuals participating in the theory's relational complexes – are the points of the spacetime manifold, which is the basic object on which fields are defined. For the spacetime structuralist, then, these points have no intrinsic properties or intrinsic identity, in accordance with ontic structural realism. Now, for the *moderate* spacetime structuralist, who adopts the view that neither objects nor relations are ontologically prior, there *are* fundamentally real spacetime points, but they are only individuated relationally, by the metric field and other key structural features of general relativity. In short, "there undoubtedly are space-time points that fulfill the function of objects[,] [b]ut instead of these objects having intrinsic properties, all there is to them is the relations in which they stand" (Esfeld and Lam 2008, 34). For a more radical structuralist, who applies the "relations only" version of ontic structural realism to the case of general relativity, there won't be anything like fundamentally real spacetime points; spacetime points will be purely emergent features of GR's relational complexes, which carry all the fundamental reality we can ascribe to the spacetime manifold. Crucially, both kinds of spacetime structuralist will emphatically deny that spacetime is purely relational, lacking anything over and above or prior to the relations between bodies. The nature of space lies between the substantival and relational extremes: it's *structural*, in the sense either that points

are real, existent individuals lacking identity independently of the relational complexes into which they enter, or the sense that points are not fundamentally real but emerge from something else that *is*, namely the relational complexes described by general relativity.

3. Leibniz's Anticipation of Spacetime Structuralism. With the structuralist view on the table, I can now launch my investigation of Leibniz, with two initial points of caution: first, showing that Leibniz was the progenitor of spacetime structuralism will necessarily involve a fair bit of interpretation and extrapolation, due to the obvious chasm between the physics of his day and the modern understanding of space and time as a unified whole described by general relativity. What I'm trying to show is that Leibniz holds a view that *in the vocabulary of his day* looks very similar to what today's spacetime structuralists say in *their* vocabulary. Second, the question of the relationship between Leibniz's ontology and his theory of phenomenal space is one of the most vexed in all of Leibniz scholarship. For the purposes of this paper, I will bracket this issue, though I think its resolution is ultimately relevant to the accuracy of the reading I advocate here. The goal of this paper is to motivate a new reading of Leibniz's theory of space taken on its own terms; I think that a serious investigation of the *Correspondence*, with these caveats in mind, will strongly suggest that my reading is correct.

Let's first look at a passage from Leibniz's third letter to Clarke, where he formulates perhaps his most famous definition of space:

> As for my own opinion, I have said more than once that I hold space to be something purely relative... I hold it to be an order of coexistences... For space denotes, in

> terms of possibility, an order of things that exist at the same time, considered as existing together, without entering into their particular manners of existing. And when many things are seen together, one consciously perceives this order of things among themselves. (2000, 14)

This passage is undoubtedly one of the sources of the canonical reading – Leibniz directly states that space is "purely relative". We ought to construe this remark, though, in light of what he says next: space is an *order* of things that exist at the same time, an order that has nothing to do with the "particular manners of existing" of its constituents. This feature of the definition is crucial; it already indicates that Leibniz thinks there's more to space than "direct" relations between bodies. It indicates, in other words, that relations between bodies are *not* direct, and *are* parasitic on something more fundamental. So even in his supposedly canonically relationist definition of space, we see hints of a more complex view. I also want to draw attention to the modal language he uses here: the spatial order has something to do with *possibility*, though the connection is unclear. I will make it more explicit soon, as it's one point on which I read Leibniz differently from the way Belot does.

Leibniz's first definition looks extremely suggestive. And what it suggests, other passages in the correspondence clarify. In his fourth letter, in response to Clarke's pleas to refine his view of space, he elaborates the view in an almost explicitly structuralist manner. I reproduce the passage in full here:

> The author contends that space does not depend on the situation of bodies. I answer: it is true, it does not depend on such or such a situation of bodies, but *it is that order*

> *which renders bodies capable of being situated, and by which they have a situation among themselves when they exist together*, as time is that order with respect to their successive position. But if there were no creatures, space and time would be only in the ideas of God. (2000, 27, emphasis mine)

Earlier, Clarke had challenged the idea that space depends on the particular arrangement of bodies; here Leibniz restates his view in light of the challenge, revealing that space in fact *does not* depend on the arrangement of bodies. He almost explicitly says that there's an underlying order, and that this underlying order itself *is* space. Space is the order that "renders bodies capable of being situated": Leibniz seems to think that there's some kind of ontologically prior relational complex, and that by virtue of taking certain places in this structure, bodies get their particular "situations". At this point we should note that the English word "situation" is a literal rendering of the Latin word "*situs*", and the concept of *situs* plays a crucial role in Leibniz's conception of space. In the *Metaphysical Foundations of Mathematics*, Leibniz defines *situs* as "mode of coexistence" and defines motion as "change of *situs* (1969, 667-668). *Situs*, in other words, is a relational property that bodies acquire by virtue of their particular place in the spatial order. Each body has a unique *situs* at any given time, given its place in the spatial order at that time; but the order that confers *situs* on bodies does not depend on the arrangement of bodies. Instead, the order *underlies* and *makes possible* the arrangement of bodies by specifying a unique but purely relational property at each place in the structure.

But what are we to make of the remark that "if there were no creatures, space and

time would be only in the ideas of God?" One might take this remark to imply that space actually *does* depend on the arrangement of bodies after all, or that Leibniz is just being inconsistent. To see that neither is the case, first recall the modal language that Leibniz uses in his first definition of space. The modal element of Leibniz's view, to my mind, connects at a fundamental level with his conception of God. To motivate the connection, consider these remarks from the *Monadology*:

> Now, since there is an infinity of possible universes in God's ideas, and since only one of them can exist, there must be a sufficient reason for God's choice, a reason which determines him toward one thing rather than another... And this is the cause of the existence of the best, which wisdom makes known to God, which his goodness makes him choose, and which his power makes him produce. (1989, 220)

On Leibniz's view of God, the latter conceives of all the possible universes and actualizes the best one. That the one he actualizes is the *best one* constitutes a "sufficient reason" for the choice to actualize it, in accordance with Leibniz's familiar dictum that there must be a sufficient reason for every event. With this view on the table, the remark about space in the mind of God makes much more sense, revealing a deep connection between space and God's creation of the world. It looks something like this: all of the possible universes exist in God's mind; the set of all possible universes includes the set of all possible spatial orders; when God actualizes the best possible universe, he also actualizes the best possible spatial order. Now, if there "were no creatures", God wouldn't yet have actualized anything; Leibniz thinks the actual world is the best possible world, and the actual world includes various and sundry

creatures. So space, considered in an abstract sense, independently of the actual spatial order, is an infinite set of *possible* structures in the mind of God.

Thus, one potentially confusing aspect of Leibniz's view turns out to be consistent with what I see as his proto-structuralism. It's not that if there were no creatures, there would be no space *because space is nothing over and above relations between bodies*; it's rather that if there were no creatures, there would be no world in the first place: by hypothesis, our world is the best possible world, and it certainly contains many creatures. And if there's no world, there's certainly no space. It seems, then, that we've cleared an important hurdle to reconstructing Leibniz's view in the way that I think it ought to be reconstructed.

We encounter another potential obstacle in a passage from his fifth letter, a passage in which he seems to propound a view at odds with what we've seen so far. Here are the relevant remarks:

> I do not say that space is an order or situation which makes things capable of being situated; that would be nonsense... I do not say, therefore, that space is an order or situation, but an order of situations, or (an order) according to which situations are disposed, and that abstract space is that order of situations when they are conceived as being possible. (2000, 61)

Leibniz here responds to Clarke's objection to the second definition of space, which I've just discussed at length. The first thing to notice is that Leibniz seems to deny directly the view of space advanced in that second definition, even seemingly declaring the earlier view to be nonsense! If this were the case, then interpretive integrity would demand that I relax my

structuralist reading. But we need to look at the way Clarke phrases his objection; in doing so, we see that he misreads Leibniz's second definition, and that Leibniz's response in this new passage is aimed at the misreading.

In his fourth reply, Clarke had objected thus: “I do not understand the meaning of these words: 'an order (or situation) which makes bodies capable of being situated'. It seems to me to amount to this: that situation is the cause of situation” (2000, 34). Notice that he *does not* object to the coherence of saying that an underlying order (structure) confers *situs* on the bodies that participate in it. He only objects to the coherence of claiming that an underlying *situation* confers *situs* on individual bodies: he thinks that it's incoherent to say that *situs* confers *situs*. Now, this claim would clearly be incoherent, but Leibniz never makes it. To see this, look back to the second definition cited above: Clarke simply inserts the parenthesis in his objection, and the parenthesis is what generates the objection in the first place. What this passage actually does, to my mind, is to reinforce the structuralist reading that I'm advocating. Leibniz agrees that *situs* can't confer *situs*, on pain of incoherence. But he never denies the claim that he had *actually made* in the second definition: the claim that an underlying spatial order is responsible for conferring *situs* on individual bodies. And in this new passage, he still holds that space is an underlying order: it's the order “according to which situations are disposed”. This remark, along with the second definition, indicates that Leibniz thinks of the spatial order as ontologically prior to the notion of *situs*: recall his assertion that space does not depend on the particular relations among bodies.

4. Spacetime Structuralism or Modal Relationism? At this point, it's hard to escape reading Leibniz as committed to space being prior to relations among bodies, in the sense that there's a deeper relational complex underwriting the latter. We've seen that *situs* is conferred upon bodies by an order that's prior to them and does not depend on them; bodies only acquire their modes of coexistence with each other by occupying places in this order. But we now might want to ask what this order really amounts to; I think I've established that it has something to do with prior spatial relations, but recently Gordon Belot has suggested that it involves a different kind of prior thing, though something that still makes Leibniz ultimately a relationist. A brief investigation and criticism of Belot's reading will help clarify my own position.

Belot argues that Leibniz holds a view close to Belot's own “modal relationism”, in the sense that Leibniz “employ[s] a notion of geometric possibility in giving content to claims about the structure of space” (2011, 173). For Belot, there are two kinds of relationists. “Conservative” relationists “identify the geometry of space with material geometry” and “give truth conditions for claims about spatial structure that differ from those of substantivalists only in quantifying over material points rather than points of space” (2011, 3). In other words, there's nothing to space prior to the relations between chunks of matter; the geometry of existent matter is the geometry of space. Relations between bodies, consequently, are direct. “Modal” relationists, by contrast, deny the identification of spatial points with material points, instead employing a kind of geometric modality, such that claims about the ultimate structure of space are about what geometric relations could *possibly* be

instantiated by *any* set of material points. For these relationists, in other words, the relations between material bodies are no longer direct, but what they're parasitic on is a kind of modal structure, rather than a set of real parts or points of physical space. The truth conditions for claims about the structure of space, then, come from the facts about geometric possibility. For example, to say that space is *finite* is to say that "there is some number N such that it is *impossible* for material points to be located more than N units away from one another"; to say that space is *infinite* is to say that there is no such number (2011, 4). And the truth of the claim that space is finite (or infinite) depends on whether there is (or is not) such a number.

Belot thinks that Leibniz holds something like the latter view, and the argument for this interpretation revolves around two claims: first, that Leibniz is clearly *not* a *conservative* relationist, since a careful reading of his remarks about space indicates that he thinks the structure of space is prior to the structure described by the actual relations between material bodies. It should be clear that I fully agree with Belot about this. Secondly, though, Belot makes the *positive* claim that the relevant texts (including some of the same passages in the *Correspondence* on which I'm relying) support the reading that the underlying structure of Leibnizian space is modal: it's an order of geometric possibility rather than any kind of prior physical order. One way to think about this is to consider the question whether Leibniz thinks "that space can profitably be thought of as composed of geometrically related parts"; Belot answers in the negative, claims that this makes Leibniz "some sort of relationist", and then argues for a modal reading of Leibniz's relationism (2011, 173). By way of illustrating *my* reading: I agree that Leibniz denies the "geometrically related parts" view, but I do not agree

that this denial makes Leibniz *any* kind of relationist; I think his view of space involves grounding the relations between material bodies on something more than a set of modal constraints on geometric relations.

The following argument will illustrate the difference between my reading and Belot's, and will also illustrate the superiority of my reading. In addition to thinking that Leibniz is a modal relationist, Belot thinks that Leibniz is committed to the structure of space being *necessary*, or the same in all possible worlds. In any possible world, for Leibniz, space is three-dimensional and Euclidean. Now, if the structure of space is the same in all possible worlds *and* is to be understood as nothing more than a network of *possible* geometric relations, then in Leibnizian terms, space must be uncreated. In other words, it must exist only in God's mind. But we've canvassed some good reasons to deny that space only exists in God's mind: this is what I take the remarks about possibility in the *Correspondence* to be getting at. In the actual world, there *is* a spatial order; this order is one of the things God actualized when he created the actual world. So Leibniz seems to think that in the actual world, space does *not* only exist in God's mind. But equally, we have good reason to deny, with Belot, that space is just a consequence of relations between bodies. So it looks like Leibniz is neither a conservative relationist nor a modal relationist.

For Leibniz, there's a sense in which *modal* relationism, when combined with the view that the structure of space is necessary, has to collapse into *conservative* relationism, since there will be nothing in the created world prior to the relations between bodies on the former combination of views. But again, Leibniz is not a conservative relationist – he thinks

that *in the created world*, the structure of space is prior to the the structure of material relations. Space *is* part of the created world after all – it doesn't only exist in God's mind – *and* space is prior to the relations between bodies. At the same time, space doesn't consist of points that have intrinsic identity; instead, space is an *order* that confers a specific property – namely, *situs* – upon bodies *in* the order, by virtue of *where* they are in the order at a particular time. This view bears a striking resemblance to spacetime structuralism.

I can now finally address the question of what the created spatial order really amounts to: is it, as the moderate structuralist thinks, a collection of fundamental relations between equally fundamental points, but such that the points have no individuality or properties except those which the relations confer upon them? Or is it, as the more radical structuralist thinks, ultimately *just* a collection of relations? Leibniz's emphatic denial, in the *Correspondence* and elsewhere, that space has anything like actual parts leads me to conclude that he conceives of the underlying spatial order as something like the more radical alternative. It's the relations that are fundamental; out of them emerges the notion of *situs*, and out of this notion in turn emerges the notion of relations between material bodies. Space only has points, or parts, in a derivative sense: fundamentally, space is an order that allows us to talk about the locations of bodies, their relative positions, and the like. Another revealing set of remarks from the *Correspondence* bolsters the suggestion that Leibniz thought of ontologically basic relations as perfectly coherent and as fundamental in his theory of space:

> As for the objection that space and time are quantities, or rather things endowed with quantity, and that situation and order are not so, I answer that order also has its

> quantity: there is in it that which goes before and that which follows; there is distance or interval. Relative things have their quantity as well as absolute ones... And therefore though time and space consist in relations, still they have their quantity. (2000, 50)

This passage, in conjunction with the other passages I've examined, suggests that Leibniz thinks of the spatial order as ultimately a set of *distance relations* that are prior to and make possible the distance relations between material bodies. Crucially, this is very similar to the situation in modern spacetime structuralism: structuralists commonly take the *metric field* to be the fundamental determinant of the structure of spacetime, though other fields play important roles; and the metric field is precisely that field which encodes spatiotemporal distance relations within the spacetime manifold.

5. Does Leibniz's Ontology Allow for a Created Spatial Order? I will conclude by noting my awareness of an issue that my reading raises in connection with Leibniz's metaphysics. I said earlier that I would bracket the problem of the relationship between Leibniz's theory of space, taken on its own terms, and his deeper metaphysical commitments, but I cannot entirely avoid it, because a tension may arise between the two in asserting that Leibniz thinks spatial relations are part of the created world. It is widely accepted that Leibniz thinks relations have only a mental, or ideal, kind of reality. Though the precise meaning of this thesis is disputed, it does imply that the spatial order, on my reading, must be ideal *and* created. The only way this is possible, in Leibnizian terms, is if the spatial order ultimately

depends on the perceptions of individual substances, or monads. One might think that this commits Leibniz to an ultimate denial of the reality of the spatial order, making the structuralist reading pointless, unless we can show that dependence on the perceptions of monads does not imply unreality for Leibniz. I believe such a solution is forthcoming in terms of the mutual coordination of the perceptions of every monad in a world. The spatial order's dependence on monadic perceptions doesn't make it “unreal” in any robust sense, for every monad's series of perceptions is coordinated with that of every other monad so as to make all the monads perceive the same publicly accessible universe – which includes the spatial order – from its point of view. In this sense, the spatial order is just as objectively real as the monads themselves, and makes possible the arrangement of bodies that each monad perceives within that order. This reading is especially plausible when we consider that the basic individuating features of Leibniz's monads are just their perceptions; any order that depends on their perceptions will only be “ideal” in a very restricted sense. It would take another paper, one devoted to Leibniz's ontology of substance, to work out these issues fully; but I believe the potential conflict can be resolved, and that the evidence I've examined in the body of this paper strongly suggests that it's worth resolving.

References

Belot, Gordon. 2011. *Geometric Possibility*. Oxford: Oxford University Press.

Clarke, Samuel and G.W. Leibniz. 2000. *Correspondence*. Ed. Roger Ariew. Indianapolis: Hackett.

Earman, John. 1989. *World Enough and Space-Time*. Cambridge, MS: MIT Press.

Esfeld, Michael and Vincent Lam. 2008. "Moderate Structural Realism about Space-Time." *Synthese* 160:27-46.

Leibniz, G.W. 1969. *Philosophical Papers and Letters.* Ed. and trans. Leroy Loemker. Dordrecht: D. Reidel.

------. 1989. *Philosophical Essays*. Trans. Roger Ariew and Daniel Garber. Indianapolis: Hackett.

Newton, Isaac. 2004. *Philosophical Writings.* Ed. Andrew Janiak. Cambridge: Cambridge University Press.

Roberts, John T. 2003. "Leibniz on Force and Absolute Motion." *Philosophy of Science* 70:553-573.

Wuthrich, Christian. 2009. "Challenging the Spacetime Structuralist." *Philosophy of Science* 76:1039-1051.

WORD COUNT: 4997

To be presented at the biennial meeting of the Philosophy of Science Association, November 2012, and then published in revised form in *Philosophy of Science*

Why do biologists use so many diagrams?

Benjamin Sheredos, Daniel C. Burnston, Adele Abrahamsen
and William Bechtel
University of California, San Diego

Abstract

Diagrams have distinctive characteristics that make them an effective medium for communicating research findings, but they are even more impressive as tools for scientific reasoning. Focusing on circadian rhythm research in biology to explore these roles, we examine diagrammatic formats that have been devised (a) to identify and illuminate circadian phenomena and (b) to develop and modify mechanistic explanations of these phenomena.

1. Prevalence and importance of diagrams in biology

If you walk into a talk and do not know beforehand whether it is a philosophy or biology talk, a glance at the speaker's slides will provide the answer. Philosophers favor text, whereas biologists shoehorn multiple images and diagrams into most of their slides. Likewise, if you attend a philosophy reading group or a biology journal club you can readily identify a major difference. Instead of verbally laying out the argument of the paper under study, the presenter in a journal club conveys hypotheses, methods, and results largely by working through diagrams from the paper. This reflects a more fundamental contrast between philosophers and biologists: their affinity for text versus diagrams is not just a matter of how they communicate once their work is done, but shapes every stage of inquiry. Whereas philosophers construct, evaluate, and revise arguments, and in doing so construct and revise sentences that convey the arguments, biologists seek to characterize phenomena in nature and to discover the mechanisms responsible for them. Diagrams are essential tools for biologists as they put forward, evaluate, and revise their accounts of phenomena and mechanisms.

Diagrams play these roles in science more generally, but we have chosen to focus on biology – in particular, on the research topic of circadian rhythms – to begin to get traction on this understudied aspect of the scientific process. Circadian rhythms are oscillations in organisms with an approximately 24-hour cycle (circa = about + dies = day). They are endogenously generated but entrained to the day-night cycle in specific locales at different times of the year. They have been identified in numerous organisms—not only animals but also plants, fungi, and even cyanobacteria—and characterize a vast array of physiological processes (e.g., basic metabolism and body temperature) and behaviors (e.g., locomotion, sleep, and responding to stimuli).

2. Diagrams and mechanistic explanation

Diagrams play a central role in biology because they are highly suited to two key tasks: (1) displaying phenomena at various levels of detail, and (2) constructing mechanistic explanations for those phenomena., Philosophers of biology have increased their attention

to those tasks over the last two decades, construing mechanisms as systems that produce a phenomenon of interest by means of the organized and coordinated operations performed by their parts (Bechtel and Richardson 1993/2010; Bechtel and Abrahamsen 2005; Machamer, Darden, and Craver 2000). To advance a mechanistic explanation, biologists must characterize the phenomenon of interest (e.g., circadian oscillations in activity), identify the mechanism they take to be responsible (e.g., a molecular "clock"), decompose it into its parts and operations, and recompose it (conceptually, physically, or mathematically) to show that the coordinated performance of these operations does indeed generate the phenomenon. Early in the discovery process scientists may identify only a few parts and operations, and hypothesize a relatively simple mechanism that can be recomposed by mentally imagining a short sequence or cycle of operations (e.g., a single gene expression feedback loop was initially posited for the molecular clock). At least in biology, further research generally uncovers additional parts and operations with complex organization and dynamics (e.g., multiple interacting feedback mechanisms constituting the overall molecular clock mechanism).

While a simple mechanistic account might be presented linguistically in the form of a narrative about how each part in succession performs its operation, diagrams generally provide particularly useful representational formats for conceptualizing and reasoning about mechanisms.[1] By displaying just a few common graphical elements in two dimensions, a diagram can visually depict a phenomenon or the organized parts and operations of an explanatory mechanism (Bechtel and Abrahamsen 2005; Perini 2005). Available elements include labels, line drawings, iconic symbols, noniconic symbols (shapes, colors), and – the device most often used for operations – various styles of arrows. The spatial arrangement of these elements can convey spatial, temporal, or functional relations that help characterize a phenomenon or mechanism. Deploying our spatial cognition on diagrams has certain advantages over language-based reasoning in constructing mechanistic explanations. Notably, scientists can mentally *animate* (Hegarty 2004) a static diagram to simulate the succession of operations by which a simple sequential mechanism produces a phenomenon. Simultaneous operations are more challenging.[2]

The primary role of diagrams for scientists is not to provide a visual format for communicating the phenomena discovered or the mechanistic accounts that explain them. Rather, diagrams of mechanisms are comparable to the plans a designer develops *before*

[1] Defining and classifying diagrams is beyond the scope of this paper; therefore, we focus on clear exemplars and set aside such formats as micrographs and animations.

[2] As researchers recognize the complicated interaction of components in a mechanism and the complex dynamics emerging from multiple simultaneous operations, they often turn to computational modeling and the tools of dynamic systems analysis to understand how the mechanism will behave, giving rise to what Bechtel and Abrahamsen (2011) characterize as *dynamic mechanistic explanations*. Jones and Wolkenhauer (in press) provide a valuable account of how diagrams contribute to the construction of such computational models. It is also worth noting that linguistic reasoning has its own advantages. We would posit that the more complex the mechanism, the more beneficial is a coordinated deployment of linguistic, diagrammatic and computational resources.

building a new machine. These are used not just to tell those actually constructing the machine how to make it; they also figure in the design process. Before producing the final plans, the designer tries out different designs and evaluates whether they are likely to result in a working and efficient machine. Often the initial sketches of these plans reveal serious problems that must be overcome, resulting in revisions to the plans. The biologist is not creating the machine (except in fields such as synthetic biology), but is trying to reverse engineer it. Still, she needs to go through many of the same processes as a designer—sketching an initial diagram, identifying ways in which it is inadequate, and modifying the diagram repeatedly until it is judged a satisfactory mechanistic account of the targeted phenomenon. Moreover, the biologist wants to end up not merely with some possible mechanism capable of producing the phenomenon, but rather with the one actually present in the biological system. In what follows, we will examine how diagrams are put to work in biology, focusing on two key tasks: delineating phenomena, and constructing mechanistic accounts to explain them.

3. Diagrams to delineate the phenomenon

An initial delineation of the phenomenon to be explained is a crucial step in mechanistic research. This remains true even if, in the course of discovering the mechanism, researchers revise their understanding of the phenomenon. Many philosophical accounts of mechanistic explanation have focused on linguistic descriptions of phenomena (e.g., "in fermentation, sugar is converted into alcohol and carbon dioxide by means of a series of intermediate reactions within yeast cells"). However, scientists focus much of their effort on obtaining much more specific, often quantitative, accounts of phenomena. Numerical data involved in characterizing a phenomenon may be presented in tables. As Bogen and Woodward (1988) made clear, however, explanations are directed not at the data but rather at the pattern extracted from the data—the phenomenon. Some data patterns can be captured in one or a few equations, such as the logarithmic function relating stimulus intensity (e.g., amplitude of a tone) to the sensation evoked (e.g., perceived loudness). By plotting these values on a graph, the phenomenon of a nonlinear relation between amplitude and loudness is immediately evident. The graph takes advantage of spatial cognition, whereas the logrithmic equation makes explicit a very precise claim that can and has been challenged (e.g., by those who argue for a power function). Scientists move deftly between linguistic descriptions, diagrams, and equations when all are available, using each to its best advantage.

Diagrams are especially useful for thinking about dynamic phenomena – patterns of change over time. Circadian phenomena are dynamic, so diagrams conveying them generally incorporate time in some way (as the abscissa on a line graph, as the order of arrows in a sketch of a mechanism, as points along the trajectory in a state space, etc.). Moreover, research on circadian oscillations often targets the interaction between endogenous control (by an internal clock) and exogenous timing cues, commonly referred to as *Zeitgebers.* Hence, what was needed was a way of diagramming the activity of an organism, such as a mouse running on a wheel, that revealed at a glance its rhythmicity and the impact of Zeitgebers.

Circadian researchers settled on a distinctive format, the *actogram*. Figure 1 illustrates the diagrammatic devices that satisfy the desiderata Time of day is represented horizontally and successive days are represented vertically (one line of data per day). Activity is tracked along each line—e.g., a single hash mark each time a mouse rotates a wheel. The bars at the top use white vs. black to represent the 24-hour light-dark conditions. Here the mouse was exposed to light from hours 4-16 during the first phase of the study (specified elsewhere as Days 1-7). During the other twelve hours of Days 1-7, and all 24 hours beginning Day 8, the mouse was kept in darkness. On Day 18, four hours after onset of activity, the mouse's rhythm was perturbed by a pulse of light. The large gray arrow directs the reader's attention to the effects of this isolated Zeitgeber. .

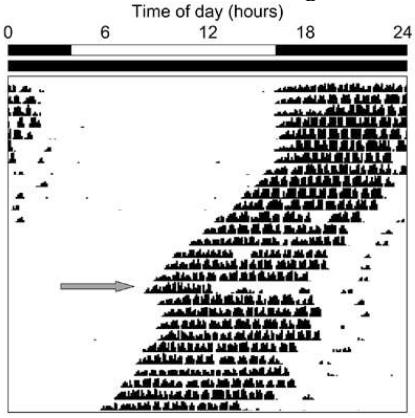


Figure 1. A basic actogram in which the top bar indicates a normal light-dark cycle for the first phase of the study (Days 1-7)_and constant darkness thereafter. The gray arrow identifies the day a light pulse was administered. (From http://www.photosensorybiology.org/id16.html.)

The actogram offers a relatively transparent representation of the animal's behavior; that is, readers who have learned its conventions should be able to see through the diagram to the multiple behavioral phenomena that it visually depicts.[3] Figure 1 offers this kind of access to at least four circadian phenomena. First, in rows 1-7 it can be seen that the hash marks occur in consolidated bands bounded by the black segments of the upper bar. This indicates that when Zeitgebers are present (light alternating with dark), virtually all wheel-running occurs in the dark: the animal is *nocturnal.* Second, the fact that the hash marks continue to appear in consolidated bands after row 7 (when the animal is free-running in the absence of Zeitgebers) indicates that the animal can endogenously maintain a robust division between periods of rest and of activity. Third, these later bands of hash marks 'drift' leftward, indicating that the animal begins its activity a bit earlier each day. Maintenance of a free-running period somewhat less than 24 hours is the core phenomenon of circadian rhythmicity. Fourth, the pulse of light flagged by the gray arrow brings an abrupt cessation of activity on Day 18 and inserts a phase delay (seen as a

[3] See Cheng (2011) for a more extensive discussion of semantic transparency. Note also that some phenomena are less transparently conveyed by diagrams than others. Presumably, the spatial cognition deployed in less transparent cases is effortful to some degree and/or coordinated with propositional cognition.

rightward "jump" in the bands of hash marks) into what was otherwise a continuing pattern of phase advance (left-ward "drift") under constant darkness. This reset phenomenon is one aspect of the more general phenomenon of *entrainment*.

Thus, actograms make circadian rhythmicity in an animal's activity visually accessible. But when chronobiologists attempt to understand the molecular mechanisms that produce such macroscopic rhythmicity, they are confronted with new phenomena that call for different diagrammatic formats. Notably, the concentration levels (relative abundance) of many types of molecules within cells oscillate. For example, Hardin, Hall, and Rosbash (1990) demonstrated the circadian oscillation of *period* (*per*) mRNA in *Drosophila melanogaster* (fruit flies).[4] In Figure 2 (below) we reproduce a pair of diagrams from their paper that illustrate how the same data can be displayed in two formats that differ substantially in how they visually depict *per* mRNA oscillation. Flies had previously been kept for three days in a light-dark cycle of 12 hours light, 12 hours dark. Starting on the fourth day (hours 24-48 in Figure 2), the flies were placed in constant darkness. Every four hours a batch of flies was sent for processing to determine *per* mRNA abundance via a molecular probe. The output of this procedure, the Northern blot, is shown at the top of Figure 2. Darker regions of the blot visually depict greater presence of *per* mRNA across the four days.

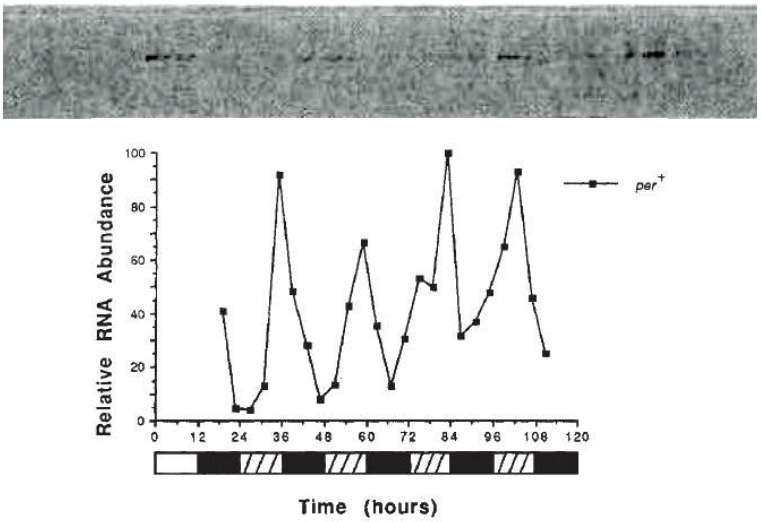


Figure 2. Two diagrams from Hardin et al.'s (1990) original portrayal of circadian oscillation in *per* mRNA levels in *Drosophila*. On top is a series of Northern blots (from different flies every 4 hours). Below this is a line graph of the same data. The Zeitgeber schedule is shown at the bottom, with white hatched bars depicting the intervals in which lights would have been on if the initial light-dark cycle had continued.

[4] Much of the early research on molecular mechanisms is nonmammalian, including the discovery of *per* mRNA oscillations. A role for *per* is conserved in the mouse circadian mechanism.

Below the Northern blots, the same data are displayed in a line graph. Here numeric values for *per* MRNA are displayed in a format that makes their oscillation immediately apparent. Moreover, a quick check of the horizontal scale confirms that the period of oscillation is circadian: there are four peaks in four days. Closer examination reveals that the peak occurs slightly earlier on Day 4, indicating a slightly shorter period in the absence of a Zeitgeber. Actograms provide a better visual display of such variations in period, but are less suitable for conveying variations in amplitude.

4. Diagrams to identify the parts, operations, and organization of a mechanism

A major use of diagrams in mechanistic science is to present a proposed mechanism by spatially displaying, at some chosen level of detail, its parts and operations and the way they are envisaged as working together to produce a phenomenon. Such diagrams typically utilize a two-dimensional space in which elements representing different parts and operations of the mechanism can be laid out so as to depict key aspects of their spatial, temporal, and functional organization. As noted in Section 2, a variety of labels, line drawings and symbols can be used to distinguish different kinds of parts. Parts perform operations that affect other parts and lead to or interact with other operations. One or more styles of arrows, often labeled, are typically chosen for displaying these operations.

As static structures, diagrams do not directly show how the mechanism produces the phenomenon. Unless a computational model is available, researchers must animate the diagram by mentally simulating the different operations and their consequences (sometimes off-loading this effort by developing animated diagrams). Such mental simulation lacks quantitative precision and can be highly fallible. A researcher may overestimate the capabilities of a component part or neglect important consequences of a particular operation, such as how it might alter another part. Moreover, diagrams themselves are generally subject to revision and quite often wrong. Since their representational content constrains what can be mentally simulated, key gaps in a diagram will yield inaccurate simulations. On the positive side, the diagram helps the researcher keep track of what must enter into each stage of simulation. In short, diagrams are an imperfect but necessary tool.

A crucial step in discovering the molecular mechanism responsible for circadian rhythms was Konopka and Benzer's (1971) discovery of *per*, the *Drosophila* gene whose mRNA levels became the focus of Hardin, Hall, and Rosbash's (1990) research. In addition to showing circadian oscillations in *per* mRNA, Hardin et al ascertained that the PER protein also oscillated with a period of approximately 24 hours but peaked several hours later than *per* mRNA. Hardin et al. recognized these oscillations as a circadian phenomenon at the molecular level, but also had the idea that *per* mRNA and PER might be parts of the mechanism that explained behavioral circadian oscillations. Combining this with their knowledge that negative feedback is a mode of organization capable of producing oscillations, they proposed three variations of a molecular mechanism whose oscillatory dynamics could be responsible for, and thereby explain, behavioral oscillations. In all three variations, PER served to inhibit *per* transcription or translation in a negative feedback loop. These are diagrammed, somewhat idiosyncratically, in Figure 3.

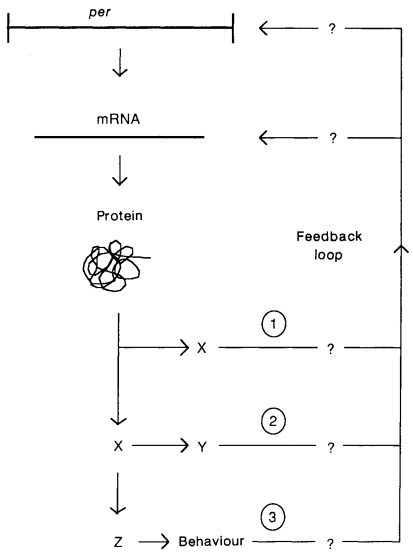


Figure 3. Hardin et al.'s (1990) representation of three versions of their proposed molecular mechanism for circadian oscillations in terms of a negative feedback loop. Question marks indicate points of uncertainty as to the origin and termination of the feedback operation.

As we claimed above, diagrams are not solely vehicles for communicating a proposed mechanistic explanation; they also can serve as a representational tool employed in reasoning about the proposed mechanism. First, a diagram can be used to envisage how a particular mechanism functions to produce a phenomenon. In this case, the phenomenon involves regular oscillations. To understand how the mechanism produces such oscillations a viewer would begin at the upper left, where the known operations of transcription into mRNA and translation into a protein are portrayed. These result in the accumulation of PER molecules, represented in the diagram as a small line drawing of one molecule. Once PER accumulates, feedback must inhibit either transcription or translation, thereby stopping the accumulation of PER. The existing PER will gradually degrade (an operation not explicitly represented, but which molecular biologists would readily infer). As it degrades, the concentration of PER will decline. This will release the transcription and translation processes from inhibition, and synthesis of PER will begin again. When repeated, this cycle of active and repressed *per* expression will result in the observed pattern of rhythmic oscillations in both *per* mRNA and PER.

A second major way in which such a diagram can serve reasoning about a mechanism is by making it clear where there are uncertainties about its operations. Note *how little* of Figure 3 is put forth as a depiction of previous discoveries concerning the mechanisms of *per* regulation. The bulk of the diagram serves as a simultaneous depiction of multiple *possible* mechanisms (sketched only in bare outline) that could explain oscillations of *per* mRNA and PER. The diagram is in large part an invitation to explanation, not a record of it. The possible mechanisms sketched here as (1) – (3) could each theoretically account for the observed oscillations. In (1), PER interacts with some biochemical substrate or process "X",

which then somehow regulates either the *per* gene itself (transcriptional regulation), or the transcribed mRNA (post-transcriptional regulation). In (2), X interacts with some further substrate or process "Y," which then does the same. In (3), the behavior of the organism provides the necessary feedback. What is known is only that the mechanism(s) at work in *Drosophila* must eventuate in regulation of *per* mRNA abundance.

Third, the constraints presented by what is presented in the diagram serve to guide hypothesizing about and investigating of further elements of the proposed mechanism. Indeed, both the unknowns represented by the question marks in Figure 3 and the operations specified became the focus of subsequent research. For example, researchers sought not merely to determine where PER fed back to inhibit formation of more PER, but how it did so. This and other inquiries quickly led to the discovery of many additional components of the mechanism: by the end of the 1990s at least seven different genes, as well as their transcripts and proteins, were viewed as part of the clock mechanism, both in *Drosophila* and in mammals. Many of these were also shown to oscillate, but at different phases than PER.

As the list of clock parts expanded and as researchers proposed multiple feedback loops, it became ever more crucial to be able to represent how the operations performed by individual parts affected other parts, and researchers regularly produced diagrams to illustrate and guide their reasoning. On the left in Figure 4 is a fairly typical contemporary diagram of the mammalian circadian oscillator. Key parts are indicated by upper-case labels: italicized for genes vs. enclosed in colored ovals for proteins. When proteins serve as transcription factors, they are shown attached to the promoter regions (E-box, D-box, and RRE) of the respective genes.

In using this diagram to reason about the mechanism, researchers follow the action of individual proteins and the ways in which they activate or repress the expression of specific genes. At the top right is a further-specified version of the feedback loop first proposed by Hardin et al. in which PER inhibits its own transcription: it does so by dimerizing with CRY (Hardin et al.'s substrate "X") and preventing the CLOCK/BMAL1 complex (Hardin et al.'s substrate "Y") from upregulating *per* transcription at the E-box promoter site. There is also a second feedback loop responsible for the synthesis of CLOCK and BMAL1. A second promoter site on the *per* gene has been identified, and its activator (DBP) is part of a positive feedback loop. It should be obvious that as the understanding of the mechanism became more complicated, diagrams became ever more crucial both in representing the mechanism and in reasoning about it. We should note that research on this mechanism is far from complete. The inhibitory operations, in particular, are the focus of important ongoing research that is serving to identify yet additional parts and operations. Diagrams such as these serve not just to represent and facilitate reasoning about the mechanism but also serve as guides to where further investigation is required (even if these are not always explicitly signaled by question marks).

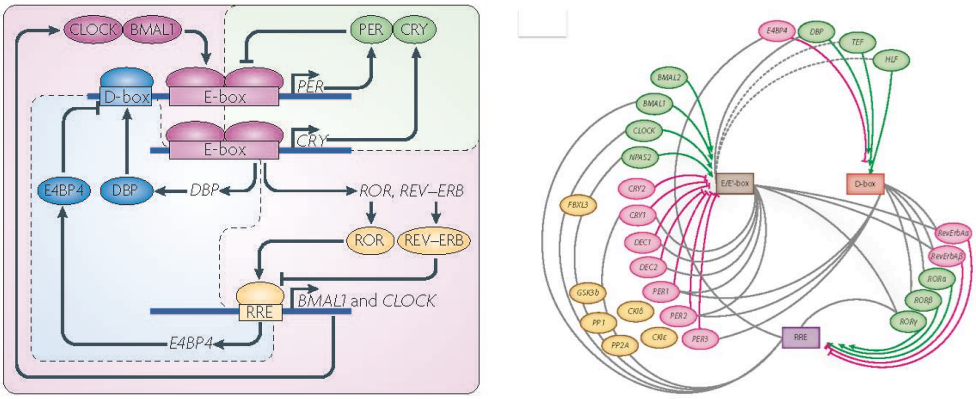


Figure 4. On the left is an example of a common way of representing the mechanism of the mammalian circadian clock, labeling genes in black italics and the proteins they express in colored ovals and using arrows to represent feedback loops (Zhang and Kay 2010). On the right an alternative representation (Ukai and Ueda 2010) which places the three promoter sites at the center. A grey line from the promoter to the gene indicates that the promoter site is found on the gene, whereas green arrows from the gene to a promoter box indicate that the protein synthesized from the gene is an activator at that promoter site and a while a squared-off magenta line indicates that the protein in some way inhibits the expression of the gene.

Once a basic diagram format is developed and researchers become familiar with its conventions, it is often retained by other researchers, who introduce relatively minor modifications to capture specific features of a given account. The choice of a diagrammatic format is not neutral, and researchers sometimes find it important to develop alternative formats that provide a different perspective on the mechanism. Ueda, for example, has introduced the alternative representation shown in the diagram on the right side of Figure 4. It presents essentially the same information about parts and operations as the diagram on the left, but shifts attention away from the genes and proteins to the promoter regions – the three boxes placed in the center of the figure. The different genes that are regulated by these promoters are shown in colored ovals in the periphery of this diagram. The proteins they express are assumed but not depicted. The relation of the boxes to the genes is explained in the figure caption.

Ueda adopted this format as part of his argument that the relations between the three promoter regions are fundamental to the functioning of the clock. Transcription factors bind to particular promoters at different times of day: the E/E' box in morning, the D box in midday, and the RRE at nighttime. For Ueda, the individual genes and proteins involved are just the vehicles via which these promoters interact. He made this even more explicit in the three diagrams shown in Figure 5. Here he abstracts from the genes and proteins and focuses just on the promoters, using arrows to indicate when products from the sites serve to activate or repress activity at another promoter. He shows all these interactions in the diagram on the left, but further decomposes them into two kinds of circuits (motifs) in the other two diagrams. In the middle is a delayed negative feedback motif in which proteins

expressed in the morning regulate expression of other genes at midday, which then repress the morning element. On the right is a repressilator motif in which products from each element repress further operation of the preceding element. Each of these motifs has been the subject of experimental, computational, and synthetic biology investigations that show how they generate oscillations (Ukai-Tadenuma et al. 2011).

Importantly, in choosing to represent the mechanism as in Figure 5, different aspects of its organization and functioning become salient. By emphasizing the overall structure of the mechanism, the overlapping oscillations are made more salient at the expense of detail about the proteins involved in the regulatory processes. These different contents provide different constraints on the reasoning that can be performed by way of the diagram, and can lead to different insights about the mechanism itself, thus helping to provide a more complete explanation of the phenomenon.

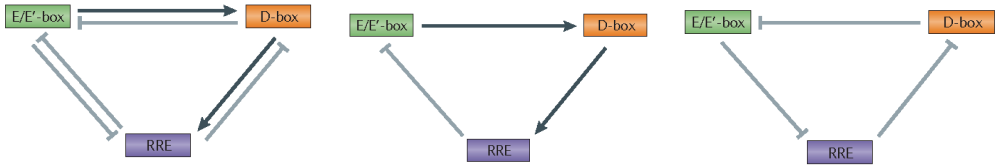


Figure 5. Hogenesch and Ueda's (2011) diagrams that abstract from the genes and proteins of the circadian oscillator to identify the basic causal circuit (left), which he then decomposes into two motifs (center and right) that are viewed as explaining the oscillatory behavior of the mechanism.

5. Conclusion: Diagrams and Mechanistic Explanation

A major explanation for the prevalence of diagrams in biology is the role they play in mechanistic explanation. We have focused on their role in two pursuits—delineating a phenomenon of interest and constructing mechanistic accounts to explain the phenomenon. A number of diagrams may be generated in making progress from an initial account to the one proposed in public. Each specifies the parts, operations, and organization of the current conception of the mechanism. Diagrams also play other roles in mechanistic explanation. For example, even modestly complex mechanisms, such as those involving negative feedback loops, challenge the ability of theorists to figure out their behavior by mentally rehearsing their interactions. To visualize dynamic phenomena, scientists often resort to other types of diagrams, such as phase spaces in which oscillations appear as limit cycles. Such diagrams abstract from mechanistic details to portray how the overall state of the system changes over time.

Having identified important roles diagrams play in biology, we conclude by noting three ways in which analysis of diagrams contributes to philosophy of science. We have begun to address the first: from diagrams we can gain a (partial) understanding of how scientists reason about a phenomenon, specifically by simulating the understood elements of a mechanism encoded in a diagram to see if they are adequate to explain the phenomenon. Second, diagrams can serve as a vehicle for understanding scientific change when we analyze how the diagrams within a field evolve, find acceptance, and are eventually

discarded. Third, identifying the cognitive elements of diagram use, including their design and the learning processes required to interpret them, can provide insight into the cognitive processes involved in scientific reasoning more generally. By directing attention to the importance of diagrams in biology, we hope to have set the stage for more sustained philosophical inquiry.

Acknowledgment

Research for this paper was support by NSF Grant 1127640.

References

Bechtel, William, and Adele Abrahamsen. (2005). "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36:421-441.

———. (2011). "Complex Biological Mechanisms: Cyclic, Oscillatory, and Autonomous," In Clifford A. Hooker, ed., *Philosophy of Complex Systems. Handbook of the Philosophy of Science*, 257-285. New York: Elsevier.

Bechtel, William, and Robert C. Richardson. (1993/2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Bogen, James, and James Woodward. (1988). "Saving the Phenomena." *Philosophical Review* 97:303-352.

Cheng, Peter C. H. (2011). "Probably Good Diagrams for Learning: Representational Epistemic Recodification of Probability Theory." *Topics in Cognitive Science* 3:475-498.

Hardin, Paul E., Jeffrey C. Hall, and Michael Rosbash. (1990). "Feedback of the *Drosophila Period* Gene Product on Circadian Cycling of Its Messenger Rna Levels." *Nature* 343:536-540.

Hegarty, Mary. (2004). "Mechanical Reasoning by Mental Simulation." *Trends in Cognitive Science* 8:280-285.

Hogenesch, John B., and Hiroki R. Ueda. (2011). "Understanding Systems-Level Properties: Timely Stories from the Study of Clocks." *Nature Reviews Genetics* 12:407-416.

Jones, Nicholaos, and Olaf Wolkenhauer. (in press). "Diagrams as Locality Aids for Explanation and Model Construction in Cell Biology." *Biology and Philosophy*.

Konopka, Ronald J., and Seymour Benzer. (1971). "Clock Mutants of *Drosophila Melanogaster*." *Proceedings of the National Academy of Sciences (USA)* 89:2112-2116.

Machamer, Peter, Lindley Darden, and Carl F. Craver. (2000). "Thinking About Mechanisms." *Philosophy of Science* 67:1-25.

Perini, Laura. (2005). "Explanation in Two Dimensions: Diagrams and Biological Explanation." *Biology and Philosophy* 20:257-269.

Ukai, Hideki, and Hiroki R. Ueda. (2010). "Systems Biology of Mammalian Circadian Clocks." *Annual Review of Physiology* 72:579-603.

Ukai-Tadenuma, Maki, Rikuhiro G. Yamada, Haiyan Xu, Jürgen A. Ripperger, Andrew C. Liu, and Hiroki R. Ueda. (2011). "Delay in Feedback Repression by Cryptochrome 1 Is Required for Circadian Clock Function." *Cell* 144:268-281.

Zhang, Eric E., and Steve A. Kay. (2010). "Clocks Not Winding Down: Unravelling Circadian Networks." *Nat Rev Mol Cell Biol* 11:764-776.

Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences

Michael Silberstein

Elizabethtown College

UMD, College Park

Anthony Chemero

University of Cincinnati

Franklin & Marshall College

Abstract. Several articles have recently appeared arguing that there really are no viable alternatives to mechanistic explanation in the biological sciences (Kaplan and Craver 2011; Kaplan and Bechtel 2011). This claim is meant to hold both in principle and in practice. The basic claim is that any explanation of a particular feature of a biological system, including dynamical explanations, must ultimately be grounded in mechanistic explanation. There are several variations on this theme, some stronger and some weaker. In order to avoid equivocation and miscommunication, in section 1 we will argue that mechanistic explanation is defined by localization and decomposition. In section 2 we will argue that systems neuroscience contains explanations that violate both localization and decomposition on any non-trivial construal of these concepts. Therefore, in section 3 we conclude the mechanistic model of explanation either needs to stretch to now include explanations wherein localization or decomposition fail, or acknowledge that there are counter-examples to mechanistic explanation in the biological sciences. We will also consider consequences and possible replies on the part of the mechanist in section 3.

1. Introduction. While there are many different accounts of mechanistic explanation, the basic idea is that a phenomenon has been explained when the responsible realizing or underlying mechanism has been identified. In particular, the relevant parts of the mechanism and the operations they perform must be identified, i.e., those parts/operations that maintain, produce, or underlie the phenomena in question (Bechtel 2010; Craver 2007; Machamer, Darden and Craver 2000; Kaplan and Craver 2011; Kaplan and Bechtel 2011). Whatever the particular account of mechanistic explanation on offer, it is clear that mechanistic explanation is supposed to be fundamental in the biological sciences, period. What is less clear is exactly what this explanatory axiom entails. What follows is a list of claims pertaining to dynamical and mathematical explanations in the biological sciences that some mechanistic thinkers assert are entailed by the mechanistic model:

1) Dynamical and mathematical explanations in systems neuroscience must be grounded in or reduced to mechanistic explanations (via localization and decomposition) to be explanatory.

2) Dynamical mechanisms are not an alternative to mechanistic explanation but a complement.

3) When dynamical and mathematical models do not describe mechanisms by appropriately mapping elements of the latter onto the former, then they provide no real explanation.

4) At this juncture, dynamical and mathematical models of explanation in biology not sufficiently grounded in mechanisms have nothing to offer but "predictivism" by way of explanatory force. That is, critics of mechanistic explanation do not have a viable alternative research strategy or alternative conception of explanation on offer (Kaplan and Bechtel 2011; Kaplan and Craver 2011).

The mechanists in question claim that certain defenders of dynamical and mathematical explanation in the biological sciences violate 1-3 and are therefore guilty of 4 (Kaplan and Craver 2011; Kaplan and Bechtel 2011). We first we need to get clear on exactly how the "dynamicist" is being portrayed. Kaplan and Craver go after the "strong dynamicist and functionalist", which they characterize as follows, "In particular, we oppose strong dynamicist and functionalist views according to which mathematical and computational models can explain a phenomenon without embracing commitments about the causal mechanisms that produce, underlie, or maintain it" (2011, 603). The strong dynamicist and functionalist holds that "mechanistic explanation is no longer an appropriate goal for cognitive and systems neuroscience" (Ibid). And finally, "If these dynamicists are right, such models yield explanations in the total absence of commitments regarding the causal mechanisms that produce the cognitive or system behavior we seek to explain" (Ibid, 604). According to Kaplan and Craver then, the strong dynamicist abandons the mechanistic model of explanation and has nothing coherent or cogent to replace it with.

We also reject strong dynamicism and functionalism so characterized. We will show however that 'either mechanistic explanation or dynamical predictivism' is a false dilemma. What we will claim is that systems biology and systems neuroscience contain robust dynamical and mathematical explanations of some phenomena in which the essential explanatory work is not be being done by localization and decomposition. More positively, the explanatory work in these models is being done by their graphical/network properties, geometric properties, or dynamical properties. We mean this claim to be true both in practice and in principle. Presumably then, what separates us from the mechanists is that they are committed to all such "higher level" explanations ultimately being discharged via localization and decomposition and we are not. However, we certainly do not think such explanations are incompatible or mutually exclusive, we have no problem calling them "complementary." Nonetheless, we will argue that graphical and dynamical

properties for example are “non-decomposable” and non-localizable features of the causal and nomological structure of the “mechanisms” in question.

We want to end this section with a sociological note of caution. A great deal of the discussion in the literature strongly suggests that what we have before us is a thinly veiled iteration of the ancient philosophical debate between competing ‘isms’ regarding the essence of mind and the essence of explanation. Take the following, “It has not escaped our attention that 3M [mechanistic model of explanation], should it be found acceptable, has dire implications for functionalist theories of cognition that are not, ultimately, beholden to details about implementing mechanisms. We count this as significant progress in thinking about the explanatory aspirations of cognitive science” (Ibid, 612). So in one corner we have the functionalist/dynamist with their usual disregard/distaste for implementing mechanisms and in the other corner the mechanist, who insists on filling in all the boxes and the equations with the really truly fundamental “causal structure.” We think that it’s time to transcend these beleaguered battle lines. That is, while we reject strong dynamicism and functionalism, and while we agree that dynamical and mechanistic explanations inevitably go hand-in-hand, we are open to the possibility that there are explanations in the biological sciences that are not best characterized in terms of localization and decomposition. To reject this possibility out of hand is as extreme as thinking that implementing mechanisms are irrelevant for explaining cognition and behavior.

When Kaplan and Craver say, “The mechanistic tradition should not be discarded lightly. After all, one of the grand achievements in the history of science has been to recognize that the diverse phenomena of our world yield to mechanistic explanation” (2011, 613), we agree. In fact, we don’t think the mechanistic tradition should be discarded. What we do think is that the mechanistic tradition understood in terms of localization and decomposition is in principle not the only effective explanatory strategy in the life sciences.

2. Counter-Examples to Localization and Decomposition in Systems Neuroscience

2.1 *Defining Localization and Decomposition.* Localization and decomposition are universally regarded as the *sine qua non* of mechanistic explanation. Identifying the parts of a mechanism and their operations necessitates decomposing the mechanism. One can use different methods to decompose a mechanism functionally, into component operations, or structurally, into component parts (Bechtel and Richardson 2010). The ultimate goal is to line up the parts with the operations they perform, this is known as localization (Ibid). Proponents of mechanistic explanation like to emphasize the way it differs from the DN-model of explanation, which is based on laws. Mechanistic explanation is not about the derivation of phenomenon from initial conditions and dynamical laws, but rather explanation via localization and decomposition.

Mechanistic explanation is reductionist in the sense that explanation is in terms of the parts of the mechanism and the operations those parts perform. Parts and operations are at a lower level of organization than the mechanism as a whole. Bechtel says that the most

conservative mechanistic account is one in which a mechanism is characterized as generating a phenomenon via a start-to-finish sequence of qualitatively characterized operations performed by identifiable component parts (2011, 534). However, Bechtel, Craver and others have recently emphasized how liberal mechanistic accounts have become. For example, Bechtel has stressed that the reductionist methodology of localization and decomposition must be "complemented" by contextualizing parts/operations both within a mechanism at a given level and between the mechanism and its environment at a higher level. The context in question includes spatial, temporal, causal, hierarchical and organizational.

We applaud and affirm the liberalization of mechanistic explanation. We assume, though, that these mechanists consider localization and decomposition as ultimately essential to mechanistic explanation. That said, we wonder what they would count as counter-examples in principle. Fortunately, Bechtel and Richardson (2010) give us some clues. They emphasize that localization and decomposition are "heuristic" strategies that sometimes fail when a system fails to be decomposable or nearly decomposable (Ibid, 13). According to them, there are two kinds of failures of decomposability or localizability: 1) when there are no component parts or operations that can be distinguished (such as a connectionist network), in which case one can only talk about organizational features—the best one can hope for here is functional decomposition, and 2) when there are component parts and operations but their individual behaviors systematically and continuously affect one another in a non-linear fashion. In this case mechanisms are not sequential but have a cyclic organization rife with oscillations, feedback loops, or recurrent connections between components. In these instances there is a high-degree of interactivity among the components and the system is non-decomposable and therefore localization will fail (Ibid, 24). In addition, if the non-linearity affecting component operations also affects the behavior of the system as a whole, such that the component properties/states are dependent on a total state-independent characterization of the system (i.e., one sufficient to determine the state and the dynamics of the system as a whole), then the behavior of the system can be called "emergent" (Ibid, 25). They emphasize that when the feedback is system wide such that almost all "The operations of component parts in the system will depend on the actual behavior and the capacities of other its components" (Ibid, 24), the following obtains. First, the behavior of the component parts considered within the system as a whole are not predicable in principle from their behavior in isolation. Second, the behavior of the system as a whole cannot be predicted even in principle from the separable Hamiltonians of the component parts (Ibid).

We affirm all this and indeed others have stressed these points in illustrating the *limits* of localization and decomposition (Chemero and Silberstein 2008; Stepp, Chemero, and Turvey 2011). However, what puzzles us is that Bechtel and Richardson go on to say that, "When these conditions are met, the systemic behavior is reasonably counted as emergent, even though it is fully explicable mechanistically" (Ibid, 24). Here Bechtel and Richardson seem to be saying that even though such "emergent" behavior is not amenable to decomposition or localization, it is nonetheless mechanistically explicable.

But, in exactly what sense are such systems *mechanistically* explicable? We shall return to this in section 3, after we consider explanations in systems neuroscience.

2.2 *Explanation in Systems Neuroscience*. Systems neuroscience is a rapidly growing area devoted to figuring out how the brain engages in the coordination and integration of distributed processes at the various length and time scales necessary for cognition and action. The assumption is that most of this coordination represents patterns of spontaneous, self-organizing, macroscopic spatiotemporal patterns which resemble the on-the-fly functional networks recruited during activity. This coordination often occurs at extremely fast time scales with short durations and rapid changes. There is a wide repertoire of models used to account for these self-orgainzing macroscopic patterns, such as oscillations, synchronization, metastability, and nonlinear dynamical coupling. Many explanatory models such as synergetics and neural dynamics combine several of these features, e.g., phase-locking among oscillations of different frequencies (Sporns 2011).

Despite the differences among these models, there are some important generalizations to be had. First, dynamic coordination is often highly distributed and non-local. Second, population coding, cooperative, or collective effects prevail. Third, time and timing is essential in a number of ways. Fourth, these processes exhibit both robustness and plasticity. Fifth, these processes are highly context and task sensitive. Regarding the third point, there is a growing consensus that such integrated processes are best viewed not as vectors of activity or neural signals, but as dynamically evolving graphs. The evidence suggests that standard neural codes such as rate codes and firing frequencies are insufficient to explain the rapid and rapidly transitioning coordination. Rather, the explanation must involve “temporal codes” or “temporal binding” such as spike timing-dependent plasticity wherein neural populations are bound by the simultaneity of firing and precise timing is essential. In these cases neurons are bound into a group or functional network as a function of synchronization in time. The key explanatory features of such models then involves various time-varying properties such as: the exact timing of a spike, the ordering or sequencing of processing events, the rich moment-to-moment context of real world activity and immediate stimulus environment, an individual’s *history* such as that related to network activation and learning, etc. All of the above can be modeled as attractor states that constrain and bias the recruitment of brain networks during active tasks and behavior (Von der Malsberg et. al, 2010).

There is now a branch of systems neuroscience devoted to the application of network theory to the brain. The formal tools of network theory are graph theory and dynamical system theory, the latter to represent network dynamics—temporally evolving dynamical processes unfolding in various kinds of networks. While these techniques can be applied at any scale of brain activity, here we will be concerned with large-scale brain networks. These relatively new to neuroscience explanatory tools (i.e., simulations) are enabled by large data sets and increased computational power. The brain is modeled as a complex system: networks of (often non-linear) interacting components such as neurons, neural assemblies and brain regions. In these models, rather than viewing the neurons, cell groups or brain regions as the basic unit of explanation, it is brain multiscale networks and their large-scale, distributed and non-local connections or interactions that are the basic unit of explanation (Sporns 2011). The study of this integrative brain function and

connectivity is primarily based in topological features (network architecture) of the network that are insensitive to, and multiply realizable with respect to, lower level neurochemical and wiring details. More specifically, a graph is a mathematical representation of some actual (in this case) biological many-bodied system. The nodes in these models represent neurons, cell populations, brain regions, etc., and the edges represent connections between the nodes. The edges can represent structural features such as synaptic pathways and other wiring diagram type features or they can represent more functional topological features such as graphical distance (as opposed to spatial distance).

Here we focus on the latter wherein the interest is in mapping the *interactions* (edges) between the local neighborhood networks, i.e., global topological features—the architecture of the brain as a whole. While there are local networks within networks, it is the global connection between these that is of greatest concern in systems neuroscience. Graph theory is replete with a zoo of different kinds of network topologies, but one of perhaps greatest interest to systems neuroscience are small-world networks as various regions of the brain and the brain as a whole are known to instantiate such a network. The key topological properties of small-world networks are: 1) a much higher clustering coefficient relative to random networks with equal numbers of nodes and edges and 2) short (topological) path length. That is, small-world networks exhibit a high degree of *topological* modularity (not to be confused with anatomical or cognitive modularity) and non-local or long-range connectivity. Keep in mind that there are many different types of small-world networks with unique properties, some with more or less *topological* modularity, higher and lower degrees (as measured by the adjacency or connection matrix), etc. (Sporns 2011; Von der Malsberg et. al 2010).

The explanatory point is that such graphical simulations allow us to *derive, predict* and *discover* a number of important things such as mappings between structural and functional features of the brain, cognitive capacities, organizational features such as degeneracy, robustness and plasticity, structural or wiring diagram features, various pathologies such as schizophrenia, autism and other "connectivity disorders" when small-world networks are disrupted, and other essential kinds of brain coordination such as neural synchronization, etc. In each case, the evidence is that the mapping between structural and topological features is at least many-one. Very different neurochemical mechanisms and wiring diagrams can instantiate the same networks and thus perform the same cognitive functions. Indeed, it is primarily the *topologica*l features of various types of small-world networks that explain essential organizational features of brains, as opposed to *lower level*, *local* purely *structural* features. Structural and topological processes occur at radically different and hard (if not impossible) to relate time-scales. The behavior and distribution of various nodes such as local networks are determined by their non-local or global connections. As Sporns puts it, "Heterogeneous, multiscale patterns of structural connectivity [small-world networks] shape the functional interactions of neural units, the spreading of activation and the appearance of synchrony and coherence" (2011, 259).

Thanks to its generality and formal power, network neuroscience has also discovered various *predictive power laws* and *scale-free invariances,* i.e., symmetry principles at work in the brain. For example, the probability of finding a node with a degree twice as

large as an arbitrary number decreases by a constant factor over the entire distribution. The explanatory power of small-world networks derives from their organizational properties, and not from the independent properties of the entities that are in small-world networks.

3. Consequences. Surprisingly, Bechtel and Richardson themselves use small-world networks as an example to illustrate that "mechanisms" of this sort require an addition to the mechanistic armament, namely, "dynamic mechanistic explanation" (Bechtel and Richardson, 2011, 16). Dynamical mechanistic explanation utilizes the tools of dynamical systems theory such as differential equations, network theory, etc., to engage in the computer simulation of complex mechanisms wherein the differential equations in question cannot be solved analytically. They claim of course that such "dynamical" explanations should nonetheless be squarely viewed as mechanistic explanation because:

> Reliance on simulations that use equations to understand the behavior of mechanisms may appear to depart from the mechanistic perspective and embrace something very much like the DN account of explanation. A simulation involves deriving values for variables at subsequent times from the equations and values at an initial time. However, simulations are crucially different from DN explanations. First, the equations are advanced not as general laws but as descriptions of the operations of specific parts of a mechanism. Second, the purpose of a computational simulation (like mental simulation in the basic mechanistic account) is not to derive the phenomenon being explained but to determine whether the proposed mechanism would exhibit the phenomenon. Finally, an important part of evaluating the adequacy of a computational model is that the parts and operations it describes are those that can be discovered through traditional techniques for decomposing mechanisms (Bechtel, 2011, 553).

There are several things that need to be said here. First, we agree that dynamical and network-type explanations are not D-N explanation and therefore cannot be guilty of "predictivism." Secondly, we agree that such explanations are nonetheless about *predicting* whether certain *causal structures* will have certain cognitive, functional or other features. Certainly, the fact that these simulations or dynamical/graphical systems predict or allow us to derive certain features does not make them explanatory. What does make them explanatory? These simulations show why certain *causal* and *nomological* structures will exhibit said features *in virtue of* their dynamical and graphical properties. Bechtel and company will balk at the word 'nomological', because the equations are not "advanced as general laws." When defending law-like explanations and the existence of laws in the special sciences, it is customary to point out that even the laws of physics do not always meet the ideals of the D-N model. That is, physical laws are often not spatiotemporally universal or free of exceptions, *ceteris paribus* clauses, idealizations and approximations. We are happpy however to forgo the word law in favor of Bechtel's phrase "organizational principles." For example, in network-based explanations the organizing principles include the aforementioned "power laws", involving self-similarity, scale-invariance and fractal patterning in space and time. Thirdly, while it may be true that one aspect of evaluating the adequacy of a computational model is that the parts and operations it describes are discovered through traditional techniques of

decomposition, it should be clear that the brain networks being described here are non-decomposable and non-localizable. There is a degree of functional decomposition for these networks but not structural decomposition. That is, localization is simply beside the point.

There is no question that graphical and dynamical simulations do describe mechanisms, but they are not merely abstract descriptions of structural mechanisms. The key question here is what's really doing the explanatory work and the answer in this case is not in the structural or lower level mechanistic details. The simulations are not merely idealizations and approximations of such lower level structural interactions. Kaplan and Craver would claim that these models are mechanistic because they meet the "3M" criterion.

> In successful explanatory models in cognitive and systems neuroscience (a) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism (2011, 611).

If what Kaplan and Craver mean to assert here is that any explanation proffered by a mathematical model of a mechanism is only truly explanatory if and only if said explanation can be reduced to or simply mapped onto the lower level structural features of the mechanism, then such mathematical models fail to be explanatory. Again, these graphical and dynamical models are non-decomposable and non-localizable. Otherwise, networks-based explanation easily meet the 3M criteria.

The key question is whether brains have the topological architectures they do in virtue of their structural mechanisms, or vice-versa? Or put another way, *in virtue of what* do graph theoretic models explain? As Bechtel himself admits, in such non-decomposable complex systems, the global topological features act as order-parameters (collective variables) that greatly constrain the behavior of the structural elements. As Sporns puts it, "a reentrant system operates less as a hierarchy and more as a heterarchy, where super- and subordinate levels are indistinct, most interactions are circular, and control is decentralized" (2011, 193). The dynamical interactions here are recurrent, recursive and reentrant. So there is no sense in which the arrow of explanation or determination is in principle exclusively from the 'lower level' structural to the 'higher' level graphical-dynamical. There is no structural, reductive or "downward-looking" explanation for the essential graphical properties of brain networks. Simply put, such *global* organizational principles or features of complex systems are not explicable in principle via localization and decomposition.

This is true for many reasons. The aforementioned many-one relationship between the structural and graphical features illustrates that specific structural features are neither necessary nor sufficient for determining global topological features. That is, topological features such as the properties of small-world networks exhibit a kind of "universality"

with respect to lower level structural details. This is why in complex systems research part of the goal is to discover power laws and other scale-invariant relations. These laws allow us to predict and explain the behavior and future time evolution of the global state of the system regardless of its structural implementation. It turns out the reason power laws are predictive and unifying is that they show *why* the macroscopic dynamics and topological features obtain across diverse lower level structural details. And the *why* has nothing to do with similar structural details of the disparate systems.

A very brief and informal characterization of universality might be helpful here. There are many cases of universality in physics at diverse scales, but the general idea is that a number of microphysically heterogeneous systems, sometimes even obeying different fundamental equations of motion, end up exhibiting the same phenomenological behavior. When this happens we say such systems share the same critical exponents and thus all belong to the same universality class. The explanandum of universality is the uniformity and convergence of large-scale behavior across many very diverse instances. That is, universality is a feature of classes of systems, not a specific system. The Renormalization Group analysis (RG) explains why specific physical systems divide into distinct universality classes in terms of the geometry or topology of the state space of systems, i.e., the so-called fixed points of the renormalization flow. Hamiltonians describing heterogeneous physical systems fall into the basin of attraction of the same renormalization group fixed point. The space of Hamiltonians contains numerous fixed points, each of which is describing different universality classes with different critical exponents and scaling functions. The microphysically diverse systems in the same universality class will exhibit a continuous phase transition, near which, their analogous macroscopic quantities will obey power laws possessing exactly the same numerical values of the critical exponents. The quantitative behavior near phase transitions exhibits this universality wherein the values of the exponents are identical.

What is interesting here is that techniques such as RG methods from statistical mechanics are being successfully applied to complex biological systems that don't have uniform parts. The occurrence of scale-invariance and hence self-similarity is the deeper reason why microphysically and mechanistically diverse systems can exhibit very similar or even identical macroscopic behavior. Thus, there is a direct route from power law behavior, scale-invariance and self-similarity to explaining why universality is true even in complex biological systems. Global topological features cannot be predicted from or derived *ab initio* from the structural features, because these are *qualitatively* different *types* of properties.

We take no position over whether these are genuine laws: we agree with Woodward (2003) that there is no need to determine whether something is a genuine law or a mere invariance to determine whether it can be used in explanation. The manner of explanation involved here is distinctly nomological. The laws found in systems neuroscience have more in common with laws found in physics than most special science laws. This is not surprising since the formal methods involved are mostly imported from physics. In fact, when it comes to the traditional virtues one expects of laws (e.g., quantifiability, universality, predictive power, satisfaction of counterfactual conditionals, explanatory

power, simplicity, unification, etc.), the laws in systems neuroscience are no worse off than most laws in physics.

Explanations in systems neuroscience are highly pluralistic involving aspects of mechanistic, dynamical, various causal and statistical-causal explanations. Many *explanatory techniques* are used in this endeavor including a host of causal and statistical modeling techniques and variety of formal/statistical measures of complexity. There are various hybrids of these explanatory patterns as well. Therefore systems neuroscience embraces *explanatory and causal pluralism* as a matter of pragmatic explanatory practice. However, the norms of such systems neuroscience explanations decidedly transcend those of localization.

Following Woodward (2003), many mechanists such as Kaplan and Craver (2011) have adopted an interventionist account of mechanistic explanation in which a mechanistic explanation is only explanatory if it allows us to manipulate various "knobs and levers" of the mechanism thereby providing us with some control over the manifestation of the phenomenon. Said control should allow us to "predict" how the system will behave if certain parts are broken, knocked-out, altered, etc. Kaplan and Craver allege that one of the things that separates dynamical explanations from real (causal) explanations, is that the former do not allow for intervention, manipulation or control. However, explanations in systems neuroscience are consistent with manipulationist or interventionist theories of explanation in general. Indeed, not just structural decompositions, but also dynamical and graphical explanations, can be and often are interventionist explanations. Mechanistic accounts of explanation that focus on localization and decomposition have no monopoly on interventionist explanation. There is nothing that says the knobs being tweaked must be structural components, they can also be global nomological features such as order-parameters or laws.

The kinds of complex biological systems under discussion here present a problem for any simplistic interventionist mechanistic model however. For example, often knock-out type experiments reveal that because of various types of plasticity, robustness/degeneracy and autonomy in complex biological systems, turning specific structural elements on or off, such as genes, has no discernable or predictable effect. In other words, we learn that such systems are non-decomposable and thus not amenable to localization. Needless to say, global organizational features such as plasticity, robustness, degeneracy and autonomy are not explicable via localization either. Therefore, very often the type of efficacious and informative manipulations one performs on such systems involves not structural components but global features such as order-parameters.

4. Conclusion. We have been arguing that the kinds of explanation common in systems neuroscience do not involve decomposition and localization. This would seem to make them non-mechanistic. It makes no difference us whether the mechanists want to stretch mechanistic explanation to include explanations wherein localization or decomposition fail, or whether they want to acknowledge that there are counter-examples to mechanistic

explanation in systems neuroscience. We do think however that these are the only options remaining to the mechanist.

We have seen that: 1) there are mathematical explanations in systems neuroscience that are not grounded in localization and decomposition in principle, 2) mathematical explanations in systems neuroscience are complementary to explanations via localization and decomposition but not reducible to them, 3) while one can sometimes map structural elements onto mathematical explanations in systems neuroscience, the mapping is at least many-one and does not allow for structural decomposition or localization and 4) systems neuroscience really does provide an explanatory alternative to localization and decomposition that greatly transcends mere "predictivism."

References

Bechtel, W. (2009). Explanation: Mechanism, Modularity and Situated Cognition, in *The Cambridge Handbook of Situated Cognition.* P. Robbins and M. Aydede (eds.). Cambridge University Press

Bechtel, W. (2010). "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science", in *Studies in History and Philosophy of Science*. A 41:321-33.

Bechtel, W. (2011). "mechanism and Biological Explanation", in *Philosophy of science*. Volume 78:4. 533-558.

Bechtel, W. Richardson, R.C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Second edition. Cambridge, MA: MIT Press/Bradford Books.

Chemero, A. Silberstein, M. (2008). "After the Philosophy of Mind: Replacing Scholasticism with Science" in *Philosophy of Science*. Volume 75, No. 1: 1-27.

Craver, C. (2007). *Explaining the Brain.* Oxford: Oxford University Press.

Craver, C. Bechtel, W. (2007). "Top-Down Causation without Top-Down Causes", in *Biology and Philosophy* 22:547-63.

Kaplan, D. Bechtel, W. (2011). "Dynamical Models: An Alternative or Complement to Mechanistic Explanations?", in *Topics in Cognitive Science* 3. 438–444.

Kaplan, D. Craver, C. (2011). "The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective", in *Philosophy of science*. Volume 78:4. 601-28.

Sporns, O. (2011). *Networks of the Brain*. MIT Press: Cambridge.

Stepp, N., A. Chemero, and M. Turvey. (2011). Philosophy for the rest of cognitive science. *Topics in Cognitive Science*, *3,* 425-437.

Von der Malsburg, C. Philips, W. Singer, W. (2010). *Dynamic Coordination and the Brain: from Neurons to Mind.* Cambridge: MIT Press.

Woodward, J. (2003). *Making Things Happen.* New York: Oxford University Press.

13

14

15

Testing a Precise Null Hypothesis: The Case of Lindley's Paradox

Jan Sprenger*

July 13, 2012

Abstract

The interpretation of tests of a point null hypothesis against an unspecified alternative is a classical and yet unresolved issue in statistical methodology. This paper approaches the problem from the perspective of Lindley's Paradox: the divergence of Bayesian and frequentist inference in hypothesis tests with large sample size. I contend that the standard approaches in both frameworks fail to resolve the paradox. As an alternative, I suggest the Bayesian Reference Criterion: (i) it targets the predictive performance of the null hypothesis in future experiments; (ii) it provides a proper decision-theoretic model for testing a point null hypothesis and (iii) it convincingly accounts for Lindley's Paradox.

*Contact information: Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

1 Introduction. Lindley's Paradox.

Lindley's Paradox is one of the most salient cases where subjective Bayesian and frequentist inference fall apart. The paradox emerges in statistical tests of point null hypotheses with high sample sizes.

Instead of starting with a theoretical definition of the paradox, we give an example with real data (Jahn, Dunne and Nelson 1987). The case at hand involved the test of a subject's claim to possess extrasensory capacities (ESP) that would enable him to affect a series of 0-1 outcomes generated by a randomly operating machine ($\theta_0 = 0.5$). The subject claimed that these capacities would make the sample mean differ significantly from 0.5.

The sequence of zeros and ones, $X_1, \ldots, X_N$, was described by a Binomial model $B(\theta, N)$. The null hypothesis asserted that the results were generated by a machine operating with a chance of $H_0 : \theta = \theta_0 = 1/2$, whereas the alternative was the unspecified hypothesis $H_1 : \theta \neq \theta_0$. The experimenters decided to observe a very long series of zeros and ones, which would give us enough evidence as to judge whether or not the null was compatible with the data.

Jahn, Dunne and Nelson (1987) report that in 104.490.000 trials, 52.263.471 ones and 52.226.529 zeros were observed. A classical, Fisherian frequentist would now calculate the z-statistic which is

$$z(x) := \sqrt{\frac{N}{\theta_0(1-\theta_0)}} \left(\frac{1}{N} \sum_{i=1}^{N} x_i - \theta_0 \right) \approx 3.61 \tag{1}$$

and reject the null hypothesis on the grounds of the very low *p-value* it induces:

$$p := P_{H_0}(|z(X)| \geq |z(x)|) \ll 0.01 \tag{2}$$

Thus, the data would be interpreted as strong evidence for extrasensory capacities. Compare this now to the result of a Bayesian analysis. Jefferys (1990) assigns a conventional positive probability $P(H_0) = \varepsilon > 0$ to the null hypothesis and calculates the *Bayes factor* in favor of the null (the ratio of prior and posterior odds):

$$B_{01}(x) := \frac{P(H_0|x)}{P(H_1|x)} \cdot \frac{P(H_1)}{P(H_0)} \approx 19$$

Hence, the data strongly favor the null over the alternative and do *not* provide evidence for the presence of ESP.

The divergence between Bayesians and frequentists can be generalized. Arguably, what is most distinctive about the above example is the large sample size. Now assume that we are comparing observation sets of different sample size N, all of which attain, in frequentist terms, the same p-value, e.g., the highly significant value of 0.01. This means that the standardized sample mean

$z(x) = \sqrt{N}(\bar{x} - \theta_0)/\sigma$ takes the same value for all observation sets, regardless of the actual sample size. However, in that case, the Bayesian evaluation of the data will become ever more inclined to the null hypothesis with increasing N. Thus, a result that speaks highly significantly against the null from a frequentist point of view can strongly support it from a Bayesian perspective. This problem has, since the seminal paper of Lindley (1957), been known as *Lindley's Paradox*.

Due to its prominence and its simplicity, Lindley's Paradox is a suitable test case for comparing various philosophies of statistical inference, and for reconsidering the goals and methods of testing a precise null hypothesis. In this paper, I ask the following questions: First, which statistical analysis of the ESP example is correct? Second, which implications has Lindley's Paradox for standard procedures of Bayesian and frequentist inference? Third, is there a full decision-theoretic framework in which point null hypothesis tests can be conducted without adopting a fully subjectivist perspective? I will argue that both the standard Bayesian and the standard frequentist way to conceive of Lindley's Paradox are unsatisfactory, and that alternatives have to be explored. In particular, I believe that José Bernardo's Bayesian Reference Criterion holds considerable promise as a replication-oriented decision model that fits our intuitions about Lindley's Paradox.

2 Testing a precise null: frequentist vs. Bayesian accounts

Lindley's Paradox deals with tests of a precise null hypothesis $H_0 : \theta = \theta_0$ against an unspecified alternative $H_1 : \theta \neq \theta_0$ for large sample sizes. But why are we actually testing a precise null hypothesis if we know in advance that this hypothesis is, in practice, never *exactly* true? (For instance, in tests for the efficacy of a medical drug, it can safely be assumed that even the most unassuming placebo will have some minimal effect, positive or negative.)

The answer is that precise null hypotheses give us a useful idealization of reality for the purpose at hand. This is also rooted in Popperian philosophy of science: "only a highly testable or improbable theory is worth testing and is actually (and not only potentially) satisfactory if it withstands severe tests" (Popper 1963, 219–220). Accepting such a theory is not understood as endorsing the theory's truth, but as choosing it as a guide for future predictions and theoretical developments.

Frequentists have taken the baton from Popper and explicated the idea of severe testing by means of statistical hypothesis tests. Their mathematical rationale is that if the discrepancy between data and null hypothesis is large

enough, we can infer the presence of a significant effect and reject the null hypothesis. For measuring the discrepancy in the data $x := (x_1, \ldots, x_N)$ with respect to postulated mean value θ_0 of a Normal model, one canonically uses the standardized statistic

$$z(x) := \frac{\sqrt{N}}{\sigma}\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0\right)$$

that we have already encountered above. Higher values of z denote a higher divergence from the null, and vice versa. Since the distribution of z usually varies with the sample size, some kind of standardization is required. Many practitioners use the *p-value* or *significance level*, that is, the "tail area" of the null hypothesis under the observed data, namely $p := P_{H_0}(|z(X)| \geq |z(x)|)$.

On that reading, a low p-value indicates evidence against the null: the chance that z would take a value at least as high as $z(x)$ is very small, if the null were indeed true. Conventionally, one says that $p < 0.05$ means significant evidence against the null, $p < 0.01$ very significant evidence, or in other words, the null hypothesis is rejected at the 0.05 level, etc. R.A. Fisher has interpreted p-values as "a measure of the rational grounds for the *disbelief* [in the null hypothesis] it augments" (Fisher 1956, 43).

Subjective Bayesians choose a completely different approach to hypothesis testing. For them, scientific inference obeys the rules of probabilistic calculus. Probabilities represent honest, subjective degrees of belief, which are updated by means of Bayesian Conditionalization. A Bayesian inference about a null hypothesis is based on the posterior probability $P(H_0|E)$, the synthesis of data E and prior $P(H_0)$.

It is here that Bayesians and significance testers clash with each other. If the p-value is supposed to indicate to what extent the null is still tenable, we get a direct conflict with Bayesian reasoning. The analyses of Berger and Delampady (1987) and Berger and Sellke (1987) show that p-values tend to grossly overstate evidence against the null, to the extent that the posterior probability of the null – and even the *minimum* of $P(H_0|x)$ under a large class of priors – is typically much higher than the observed p-value. In other words, even a Bayesian analysis that is maximally biased against the null is still less biased than a p-value analysis. This has led Bayesian statisticians to conclude that "almost anything will give a better indication of the evidence provided by the data against H_0" (Berger and Delampady 1987, 330). These findings are confirmed by methodologists in the sciences who have repeatedly complained about the illogic of p-values (and significance testing) and their inability to answer the questions that really matter for science (Cohen 1994; Royall 1997; Goodman 1999).

Lindley's Paradox augments this divergence of a Bayesian and a frequentist analysis. In a Normal model, if $P(H_0) > 0$ and $N \to \infty$, then the posterior probability of the null $P(H_0|x)$ converges to 1 for almost any prior distribution over H_1. More precisely:

> **Lindley's Paradox:** Take a Normal model $N(\theta, \sigma^2)$ with known variance σ^2, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, assume $P(H_0) > 0$ and any regular proper prior distribution on $\{\theta \neq \theta_0\}$. Then, for any testing level $\alpha \in [0, 1]$, we can find a sample size $N(\alpha)$ and independent, identically distributed data $x = (x_1, \ldots, x_N)$ such that
>
> 1. The sample mean $\bar{x}$ is significantly different from θ_0 at level α;
> 2. $P(H_0|x)$, that is, the posterior probability that $\theta = \theta_0$, is at least as big as $1 - \alpha$. Lindley (cf. 1957, 187)

One might conjecture that this Bayesian-frequentist divergence stems from the unrealistic assumption that $P(H_0) > 0$. But actually, the findings are confirmed if we switch to an analysis in terms of Bayes factors, the Bayesian's standard measure of evidence. The evidence x provides for H_0 vis-à-vis H_1 is written as B_{01} and defined as the ratio of prior and posterior odds:

$$B_{01}(x) := \frac{P(H_0|x)}{P(H_1|x)} \cdot \frac{P(H_1)}{P(H_0)} = \frac{P(x|H_0)}{P(x|H_1)}, \tag{3}$$

which can alternatively be interpreted as an averaged likelihood ratio of H_0 vs. H_1. Now, if the prior over H_1, that is, the relative weight of alternatives to the null, follows a $N(\theta_0, \tilde{\sigma}^2)$-distribution, then the Bayes factor in favor of the null can be computed as

$$B_{01}(x) \quad = \quad \sqrt{1 + \frac{N\tilde{\sigma}^2}{\sigma^2}}\, e^{\frac{-Nz(x)^2}{2N + 2\sigma^2/\tilde{\sigma}^2}}, \tag{4}$$

which converges, for increasing N, to infinity as the second factor is bounded (Bernardo 1999, 102). This demonstrates that the precise value of $P(H_0)$ is immaterial for the outcome of the subjective Bayesian analysis.

This result remarkably diverges from the frequentist finding of significant evidence against the null. What has happened? If the p-value, and consequently the value of $z(X) = c$, remain constant for increasing N, we can make use of the Central Limit Theorem: $z(X)$ converges, for all underlying distributions with bounded second moments, in distribution against $N(0, 1)$. Thus, as $N \to \infty$, we obtain that $c\sigma \approx \sqrt{N}(\bar{X} - \theta_0)$, and $\bar{X} \to \theta_0$. In other words, the sample mean gets ever closer to θ_0, favoring the null over the alternatives. For the deviance between the variance-corrected sample mean z and H_0 will be relatively small compared to the deviance between z and all those hypotheses in H_1 that are

"out there", in sharp contrast to a frequentist tester who will observe significant evidence against H_0.

In other words: as soon as we take our priors over H_1 seriously, as an expression of our uncertainty about which alternatives to H_0 are more likely than others, we will, in the long run, end up with results favoring θ_0 over an unspecified alternative. Bayesians read this as the fatal blow for frequentist inference since an ever smaller deviance of the sample mean $\bar{x}$ from the parameter value θ_0 will suffice for a highly significant result. Obviously, this makes no scientific sense. Small, uncontrollable biases will be present in any record of data, and frequentist hypothesis tests are unable to distinguish between *statistical significance* ($p < 0.05$) and *scientific significance* (a real effect is present). A Bayesian analysis, on the other hand, accounts for this insight: as $\bar{X} \to \theta_0$, an ever greater chunk of the alternative H_1 will diverge from $\bar{X}$, favoring the null hypothesis.

Still, the subjective Bayesian stance on hypothesis tests leaves us with an uneasy feeling. Assigning a strictly positive degree of belief $P(H_0) > 0$ to the point null hypothesis $\theta = \theta_0$ is a misleading and inaccurate representation of our subjective uncertainty. In terms of degrees of belief, θ_0 is not that different from any value $\theta_0 \pm \varepsilon$ in its neighborhood. Standardly, we would assign a continuous prior over the real line, and there is no reason why a set of measure zero, namely $\{\theta = \theta_0\}$, should have a strictly positive probability. But if we set $P(H_0) = 0$, then for most priors (e.g., an improper uniform prior) the posterior probability distribution will not peak at the null value, but somewhere else. Thus, the apparently innocuous assumption $P(H_0) > 0$ has a marked impact on the result of the Bayesian analysis.

A natural reply to this objection contends that H_0 is actually an idealization of the hypothesis $|\theta - \theta_0| < \epsilon$, for some small ϵ, rather than a precise point null hypothesis $\theta = \theta_0$. Then, it would make sense to use strictly positive priors. Indeed, it has been shown that point null hypothesis tests in terms of Bayes factors approximate a test of whether a small interval around the null contains the true parameter value (Theorem 1 in Berger and Delampady 1987). Seen that way, it *does* make sense to assign a strictly positive prior to H_0.

Unfortunately, this won't help us in the situation of Lindley's Paradox: when $N \to \infty$, the convergence results break down, and testing a point null is no more analogous to testing whether a narrow interval contains θ. In the asymptotic limit, the Bayesian cannot justify the strictly positive probability of H_0 as an approximation to testing the hypothesis that the parameter value is close to θ_0 – which is the hypothesis of real scientific interest. Setting $P(H_0) > 0$ may be regarded as a useful convention, but this move neglects that a hypothesis test in science asks, in the first place, if H_0 is a reasonable simplification of a more general model, and not if we assign a high degree of belief to this precise value

of θ.

This fact may be the real challenge posed by Lindley's Paradox. In the debate with frequentists, the Bayesian likes to appeal to "foundations", but working with strictly positive probabilities of the null hypothesis is hard to justify from a foundational perspective, and also from the perspective of scientific practice.

The bottom line of all is that the subjective Bayesian analysis fails to explain why hypothesis tests have such an appeal to scientific practitioners, and even to those that are statistically sophisticated. Similarly, the Bayesian has a hard time to explain why informative and precise, but improbable hypotheses should sometimes be preferred over more general alternatives. How can the subjectivist model that we are less interested in the *truth* of H_0 than in its *usefulness*?

3 The BRC approach to hypothesis testing

This section presents a proposal for a fully Bayesian decision model for hypothesis testing that survives the criticisms raised against the subjectivist approach and gives a satisfactory treatment of Lindley's Paradox. The main idea is to decouple the idea of testing a precise null hypothesis H_0 from the truth of this hypothesis. Instead, we view the statistical test as making a decision on whether or not we should treat the null hypothesis $H_0 : \theta = \theta_0$ as a proxy for the more general model $H_1 : \theta \neq \theta_0$. In other words, we test whether the null is compatible with the data using a specific utility structure, going back to the roots of Bayesianism in decision theory.

Thus, we have to extend Bayesian belief revision to Bayesian decision models and add a proper utility dimension. This allows for much more flexible treatments than the traditional zero-one loss model that is implicitly presupposed in inference to the most probable hypothesis. In the remainder, I sketch a simplified version of Bernardo's Reference Bayesian Criterion (1999, section 2-3) in order to elaborate the main ideas of philosophical interest.

In science, we generally prefer hypotheses on whose predictions we may rely. Therefore, a central component of the envisioned decision model depends on the expected predictive accuracy of the null. Hence, we need a function that evaluates the predictive score of a hypothesis, given some data y. The canonical approach consists in the logarithmic score $\log P(y|\theta)$ (Good 1952): if an event considered to be likely occurs, then the score is high; if an unlikely event occurs, the score is low. This is a natural way of rewarding good and punishing bad predictions.

A generalization of this utility function describes the score of data y under parameter value θ as $q(\theta, y) = \alpha \log P(y|\theta) + \beta(y)$, where α is a scaling term, and

$\beta(y)$ is a function that depends on the data only. Informally speaking, $q(\cdot, \cdot)$ is decomposed into a prediction-term and a term that depends on the desirability of an outcome, where the latter will eventually turn out to be irrelevant. This is a useful generalization of the logarithmic score. Consequently, if θ is the true parameter value, the utility of taking H_0 as a proxy for the more general model H_1 is

$$\int q(\theta_0, Y)\, dP_{Y|\theta} \quad = \quad \alpha \int \log P(y|\theta_0)\, P(y|\theta)\, dy + \int \beta(y)\, P(y|\theta)\, dy.$$

The overall utility U of a decision, however, should not only depend on the predictive score, as captured in q, but also on the cost c_j of selecting a specific hypothesis H_j. Ceteris paribus, H_0 should be preferred to H_1 because it is more informative, simpler, and less prone to the risk of overfitting (in case there are nuisance parameters). Therefore it is fair to set $c_1 > c_0$. Writing $U(\cdot, \theta) = \int q(\cdot, Y)\, dP_{Y|\theta} - c_j$, we then obtain

$$U(H_0, \theta) = \alpha \int \log P(y|\theta_0)\, P(y|\theta)\, dy + \int \beta(y)\, P(y|\theta) dy - c_0$$

$$U(H_1, \theta) = \alpha \int \log P(y|\theta)\, P(y|\theta)\, dy + \int \beta(y)\, P(y|\theta) dy - c_1.$$

Note that the utility of accepting H_0 is evaluated against the true parameter value θ, and that the alternative is not represented by a probabilistic average (e.g., the posterior mean), but by its best element, namely θ. This is arguably more faithful than subjective Bayesianism to the essential asymmetry in testing a point null hypothesis. Consequently, the difference in *expected utility*, conditional on the posterior density of θ, can be written as

$$\begin{aligned}
& \int_{\theta \in \Theta} \left(U(H_1, \theta) - U(H_0, \theta)\right)\, P(\theta|x)\, d\theta \\
= \quad & \alpha \int_{\theta \in \Theta} \left(\int \log \frac{P(y|\theta)}{P(y|\theta_0)}\, P(y|\theta) \right)\, P(\theta|x)\, dy\, d\theta + \int \beta(y)\, P(y|\theta)\, dy \\
& - \int \beta(y)\, P(y|\theta)\, dy + c_0 - c_1 \\
= \quad & \alpha \int_{\theta \in \Theta} \left(\int \log \frac{P(y|\theta)}{P(y|\theta_0)}\, P(y|\theta)\, dy \right) P(\theta|x)\, d\theta + c_0 - c_1.
\end{aligned}$$

This means that the expected utility difference between inferring to the null hypothesis and keeping the general model is essentially a function of the expected log-likelihood ratio between the null hypothesis and the true model, calibrated against a "utility constant" $d^*(c_0 - c_1)$. For the latter, Bernardo suggests a conventional choice that recovers the well-probed scientific practice of regarding three standard deviations as strong evidence against the null. The exact value

of d^* depends, of course, on the context: on how much divergence is required to balance the advantages of working with a simpler, more informative, and more accessible model (Bernardo 1999, 108).

Wrapping up all this, we will reject the null if and only if $\mathbb{E}_\theta[U(H_1, \theta)] > \mathbb{E}_\theta[U(H_0, \theta)]$ which amounts to the

> **Bayesian Reference Criterion (BRC):** Data x are incompatible with the null hypothesis $H_0 : \theta = \theta_0$, assuming that they have been generated from the probability model $(P(\cdot|\theta), \theta \in \Theta)$, if and only if
>
> $$\int_{\theta \in \Theta} P(\theta|x) \left(\int \log \frac{P(y|\theta)}{P(y|\theta_0)} \, P(y|\theta) \, dy \right) d\theta > d^*(c_0 - c_1). \qquad (5)$$

This approach has a variety of remarkable features. First, it puts hypothesis testing on firm decision-theoretic grounds. Second, accepting the null, that is, using θ_0 as a proxy for θ, amounts to claiming that the difference in expected predictive success of θ_0 and the true parameter value θ will be offset by the fact that H_0 is more elegant, more informative and easier to test. Hence, BRC does not only establish a tradeoff between different epistemic virtues: it is also in significant agreement with Popper's view that "science does not aim, primarily, at high probabilities. It aims at high informative content, well backed by experience." (Popper 1934/59, 399). Third, the approach is better equipped than subjective Bayesianism to account for frequentist intuitions, since under some conditions, e.g., in Lindley's Paradox, the results of a reference Bayesian analysis agree with the results of a frequentist analysis, as we shall see below. Fourth, it is invariant of the particular parametrization, that is, the final inference does not depend on whether we work with θ or a 1:1-transformation $g(\theta)$. Fifth, it is neutral with respect to the kind of prior probabilities that are fed into the analysis.[1]

4 Revisiting Lindley's Paradox

We now investigate how Bernardo's approach deals with Lindley's Paradox and return to the ESP example from the introduction. It turns out that the BRC quantifies the expected loss from using θ_0 as a proxy for the true value θ as substantial. Using a $\beta(1/2, 1/2)$ reference prior for θ (Bernardo 1979), the expected loss under the null hypothesis is calculated as $d(\theta = 1/2) \approx \log 1400 \approx 7.24$. This establishes that "under the accepted conditions, the precise value $\theta_0 = 1/2$ is rather incompatible with the data" (Bernardo 2012, 18).

[1] BRC implies that some parameters, such as d^*, which have to be chosen conventionally or context-dependent. Hence, a charge of "arbitrariness" could be made. However, this flexibility is, in my opinion, an asset of a general decision-theoretic model, not a drawback, as a comparison with Expected Utility Theory makes clear.

We observe that the results of a reference analysis according to BRC agree with the results of the frequentist analysis, but contradict the subjective Bayesian results. One might thus object that Bernardo's Bayesianism is purely *instrumental*: that is, it makes use of Bayesian notation and assigns a "probability" over θ, but it ends up with conventional, automated inference procedures that recover frequentist results.

Let us get back to the experiment. Of course, the rejection of the null hypothesis does not prove the extrasensory capacities of our subject; a much more plausible explanation is a small bias in the random generator. This is actually substantiated by looking at the posterior distribution of θ: due to the huge sample size, we find that for any non-extreme prior probability function, we obtain the posterior $\theta \sim N(0.50018, 0.000049)$, which shows that most of the posterior mass is concentrated in a narrow interval that does *not* contain the null. These findings agree with a likelihood ratio analysis: if we compute the log-likelihood ratio $L_{\hat{\theta},\theta_0}$ of the maximum likelihood estimate $\hat{\theta}(x_1, \ldots, x_n) = \bar{x}$ versus the null, we obtain (using the Normal approximation)

$$\begin{aligned} \log L_{\hat{\theta},\theta_0}(x_1, ..., x_N) &= \log \frac{P(\bar{x}|\hat{\theta})}{P(\bar{x}|\theta_0)} = \log \frac{P(x_1 = \hat{\theta}|\hat{\theta})^N}{P(x_1 = \hat{\theta}|\theta_0)^N} \\ &= \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N - \log \left(\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-\frac{N}{2\sigma^2}(\hat{\theta}-\theta_0)^2} \right) \\ &= \frac{N}{2\sigma^2}(\hat{\theta} - \theta_0)^2 \overset{N\to\infty}{\longrightarrow} \infty. \end{aligned} \tag{6}$$

This analysis clearly shows that the likelihood ratio with respect to the maximum likelihood estimate speaks, for large N, increasingly *against* the null (in our case: $\log L_{\hat{\theta},\theta}(x_1, ..., x_N) \approx 6.53$), in striking disagreement with the Bayes factor analysis.

If we revisit Jeffery's analysis in the light of these observations, we note two contentious features, already touched upon previously. The first concerns the utility structure that is imposed by basing inference exclusively on the posterior distribution. We have seen in the previous sections that such a zero-one loss function, and a positive prior probability $P(H_0)$ may not be adequate assumptions for deciding whether a hypothesis should be judged as compatible with the data; therefore we should also be wary of judgments based on such assumptions. Second, a Bayes factor comparison effectively compares the likelihood of the data under H_0 to the *averaged likelihood* of the data under H_1. However, this quantity is strongly influenced by whether there are some extreme hypotheses in H_1 that fit the data poorly. Compared to the huge amount of data that we have just collected, the impact of these hypotheses (mediated via the conventional uniform prior) should be minute. These arguments explain why most

people would tend to judge the data as incompatible with the *precise* null, but fail to see a scientifically interesting effect.

From the vantage point of whether the experimental effect is likely to be *replicated* – and this is a question scientists are definitely interested in – the BRC approach is more adequate. After all, it focuses on expected future success, and not on past performance. H_0 is not accepted because it is considered likely to be true, but because it is sufficiently likely to be *predictively successful*.

Frequentists may object that Bernardo's approach is a very complicated way to obtain a simple result. After all, if we use *confidence intervals* instead of p-values, we will be able to appreciate the small effect size as well as the fact that the data are incompatible with the null hypothesis. A similar point can be made in Mayo's (1996) error-statistical framework: only a small discrepancy from the null hypothesis is warranted with a high degree of severity, but no discrepancy that points to a substantial extrasensory influence rather than to a tiny bias in the machine. Hence, Lindley's Paradox seems to vanish in thin air if we only adopt the right frequentist perspective.

To this point I have a twofold reply: First, confidence intervals and severity functions are, on a mathematical level, intimately connected to p-values and Neyman-Pearson error probabilities. Therefore they share a lot of the foundational problems of p-values, some of which have been mentioned above (see Royall 1997, for an elaborate discussion). A fully convincing reply to these criticisms is still pending. Second, confidence intervals do not involve a decision-theoretic component; they are interval estimators. They do not determine whether a precise null hypothesis should be accepted or rejected. (The case is a bit more complicated for Mayo's error statistics, but as I understand her, the kind of inferences she wants to make is about *severely warranted discrepancies from the null*, and not about decisions to accept or to reject a point null hypothesis.) If we take statistical tests to be serious decision problems, if the word "test" is more than a dummy for our preferred inference problem, then those frequentist techniques do not provide a convincing account of hypothesis testing.

5 Conclusions

We have demonstrated how Lindley's Paradox – the extreme divergence of Bayesian and frequentist inference in tests of a precise null hypothesis with large sample size – challenges the standard methods of both Bayesian and frequentist inference. Neither frequentist significance tests nor subjective Bayesian inference provides a convincing account of the problem. Therefore, I have introduced Bernardo's Bayesian Reference Criterion (BRC) as a full Bayesian, albeit not subjectivist model of testing a precise null hypothesis. It turns out that

BRC gives a sensible Bayesian treatment of Lindley's Paradox, with a focus on predictive performance and likely replication of the effect in deciding whether to accept or to reject the null. The motivation of BRC also exhibits a notable similarity to ideas voiced by Karl Popper.

Of course, Bernardo's reference Bayesian approach is not immune to objections. But anyway, BRC underlines that Bayesian inference in science need not necessarily infer to highly probable models – a misconception that is perpetuated in post-Carnapian primers on Bayesian inference and that has attracted Popper's understandable criticism. To provide some evidence: Howson and Urbach (1993, xvii) claim that "scientific reasoning is essentially reasoning in accordance with the formal principles of probability" and Earman (1992, 33) even takes, in his exposition of Bayesian reasoning, the liberty of announcing that "issues in Bayesian decision theory will be ignored". As argued in the paper, such a purely probabilistic Bayesianism falls short of an appropriate model of scientific reasoning.

In other words, Bayesianism should not be separated from its decision-theoretic component that involves, beside the well-known probabilistic representation of uncertainty, also a utility function of equal significance. Failure to appreciate this fact is, to my mind, partly responsible for the gap between the debates in statistical methodology and confirmation theory. This paper makes an attempt to bridge it.

References

Berger, J.O., and M. Delampady (1987): "Testing Precise Hypotheses", *Statistical Science* 2, 317–352.

Berger, J.O., and T. Sellke (1987): "Testing a point null hypothesis: The irreconciliability of P-values and evidence", *Journal of the American Statistical Association* 82, 112–139.

Bernardo, J.M. (1979): "Reference posterior distributions for Bayesian inference", *Journal of the Royal Statistical Society B* 41, 113–147.

Bernardo, J.M. (1999): "Nested Hypothesis Testing: The Bayesian Reference Criterion", in J. Bernardo et al. (eds.): *Bayesian Statistics 6: Proceedings of the Sixth Valencia Meeting*, 101–130. Oxford: Oxford University Press.

Bernardo, J.M. (2012): "Integrated objective Bayesian estimation and hypothesis testing", in J.M. Bernardo et al. (eds.): *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*, 1–68. Oxford: Oxford University Press.

Cohen, J. (1994): "The Earth is Round ($p < .05$)", *American Psychologist* 49, 997-1001.

Earman, J. (1992): *Bayes or Bust?*. Cambridge/MA: The MIT Press.

Fisher, R.A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.

Good, I.J. (1952): "Rational Decisions", *Journal of the Royal Statistical Society B* 14, 107–114.

Goodman, S.N. (1999): "Towards Evidence-Based Medical Statistics. 1: The P Value Fallacy", *Annals of Internal Medicine* 130, 1005–1013.

Howson, C. and P. Urbach (1993): *Scientific Reasoning: The Bayesian Approach*. Second Edition. La Salle: Open Court.

Jahn, R.G., B.J. Dunne and R.D. Nelson (1987): "Engineering anomalies research", *Journal of Scientific Exploration* 1, 21–50.

Jefferys, W.H. (1990): "Bayesian Analysis of Random Event Generator Data", *Journal of Scientific Exploration* 4, 153–169.

Lindley, D.V. (1957): "A statistical paradox", *Biometrika* 44, 187–192.

Mayo, D.G. (1996): *Error and the Growth of Experimental Knowledge*. Chicago & London: The University of Chicago Press.

Popper, K.R. (1934/59): *Logik der Forschung*. Berlin: Akademie Verlag. English translation as *The Logic of Scientific Discovery*. New York: Basic Books, 1959.

Popper, K.R. (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Harper.

Royall, R. (1997): *Scientific Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Geodesic Universality in General Relativity*

Michael Tamir

Abstract

According to (Tamir, 2012), the geodesic principle strictly interpreted is compatible with Einstein's field equations only in pathologically unstable circumstances and, hence, cannot play a fundamental role in the theory. In this paper it is shown that geodesic dynamics can still be coherently reinterpreted within contemporary relativity theory as a universality thesis. By developing an analysis of universality in physics, we argue that the widespread geodesic clustering of diverse free-fall massive bodies observed in nature qualifies as a universality phenomenon. We then show how this near-geodetic clustering can be explained despite the pathologies associated with strict geodesic motion in Einstein's theory.

1 Introduction

In Einstein's original conception of the general theory of relativity, the behavior of gravitating bodies was determined by two laws: The first (more fundamental) law consisted of his celebrated field equations describing how the geometry of spacetime is influenced by the flow of matter-energy. The second governing principle, referred to as the *geodesic principle*, then provides the "law of motion" for how a gravitating body will "surf the geometric field" as it moves through spacetime. According to this principle a gravitating body traces

*Thanks to John Norton, Robert Batterman, and Balázs Gyenis for many helpful conversations.

out the "straightest possible" or *geodesic* paths of the spacetime geometry. Not long after the theory's initial introduction, it became apparent that the independent postulation of the geodesic principle to provide the theory's law of motion was redundant. In contrast to classical electrodynamics and Newtonian gravitation, general relativity seemed special in that its dynamics providing principle could be derived directly from the field equations.

Though the motion of gravitating bodies is not logically independent of Einstein's field equations, the geodesic principle canonically interpreted as providing a precise prescription for the dynamical evolution of massive bodies in general relativity does not follow from Einstein's field equations. To the contrary, in (Tamir, 2012) it was argued that under the canonical interpretation, *not only does the geodesic principle fail to follow from the field equations, but such exactly geodetic evolution would generically violate the field equations for non-vanishing massive bodies*. In short, under the canonical interpretation the two laws are not even consistent.

Despite this failure, the widespread "approximately geodetic" motion of free-fall bodies must not be denied. The nearly-geodetic evolution of gravitating bodies is well confirmed within certain margins of error. Moreover, some of the most important confirmations of Einstein's theory, including the classic recovery of the otherwise anomalous perihelion of Mercury, also appear to confirm the approximately geodetic motion of massive bodies. This abundance of apparent confirmation suggests that though the claim that massive bodies must exactly follow geodesics fails to cohere with Einstein's theory, geodesic following may constitute some kind of *idealization* or *approximately correct* description of how generic massive bodies behave.

We must hence reconcile an apparent dilemma: On the one hand geodesic following appears illustrative as an ideal of the true motion of massive bodies. On the other hand the arguments against the canonical view in (Tamir, 2012) reveal that non-vanishing bodies that actually follow geodesics would be highly pathological with respect to the theory, suggesting that they are not suitable as ideal theoretical models. Moreover, even if we were to adopt such models as idealizations, in order to gain knowledge about the paths of *actual* bodies, it is unclear how to draw conclusions about the non-pathological cases by considering pathological models that are generically incompatible with the theory.

In this paper, we establish such a reconciliation by arguing that, in light of the failure of the canonical interpretation, the principle should instead be adopted as a *universality thesis* about the clustering of certain classes of gravitating bodies that exhibit *nearly*-geodetic motion. In section 2, we propose an analysis of the general concept of *universality phenomena* to designate a certain kind of similarity of behavior exhibited across a wide class of (ostensibly diverse) systems of a particular theory. Using this analysis, in section 3, we explain how the nearly geodetic behavior observed in numerous gravitational systems counts as such a clustering within appropriately close (topological) neighborhoods of *anchor models* that exhibit perfect geodesic motion. Finally, in section 4, we explain why such pathological anchor models can be employed to characterize this clustering of the realistic models, without having to reify the problem models or take them as representative of actual physical systems.

2 Universality in Physics

The arguments of (Tamir, 2012) reveal that the geodesic principle cannot be used to prescribe the precise dynamics of massive bodies in general relativity. Nevertheless, the geodesic principle, demoted from the status of fundamental law to a thesis about the general motion of classes of gravitating bodies, may still be of value to our understanding generic dynamical behavior in general relativity. The challenge is to find an appropriate way of characterizing such "nearly geodetic" motion in terms of closeness to perfect geodesic following motion in light of the fact that attempts to model gravitating bodies that could stably follow geodesics end up violating Einstein's field equations. If such a reinterpretation of the principle is well-founded, we must justify its endorsement in the face of the kinds of pathologies associated with actual geodesic motion. This can be done by interpreting the robust geodesic clustering patterns actually observed in nature as a *universality phenomena*. In this section, we begin with an explicit analysis of this concept's use in physics.

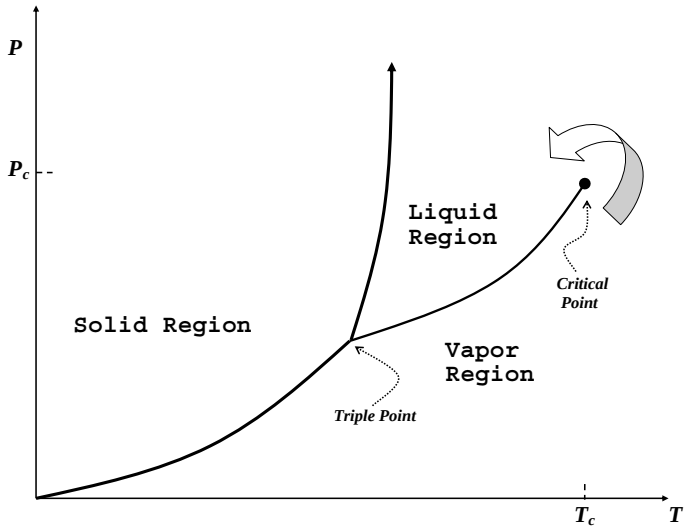


Figure 2.1: *Phase diagram of a generic material at fixed density.*

2.1 The Paradigm Case: Universality in Phase Transitions

The notion of a universality phenomenon was initially coined to characterize a remarkable clustering in the behavior of thermal systems undergoing phase transitions, particularly the behavior of systems in the vicinity of a thermodynamic state called the "critical point." In thermodynamics the state of a system can be characterized by the three state variables pressure, temperature, and density. According to the thermodynamic study of phase transitions, when the state of a system is kept below the particular "critical point" values (P_c, T_c, ρ_c) associated with the substance, phase transition boundaries correspond to discrete changes in the system (signified in figure 2.1 by the thick black lines). If, however, a system is allowed to exceed its critical values, there exist paths available to the system allowing it to change from vapor to liquid (or back) without undergoing such discrete changes. These paths involve avoiding the vapor-liquid boundary line by navigating around the critical point as depicted by the broad arrow in figure 2.1.

There exists a remarkable uniformity in the behavior of different systems near the critical point. One such uniformity is depicted in figure 2.2. In this figure we see a plot of data recovered by Guggenheim (1945) in a temperature-density graph of the thermodynamic

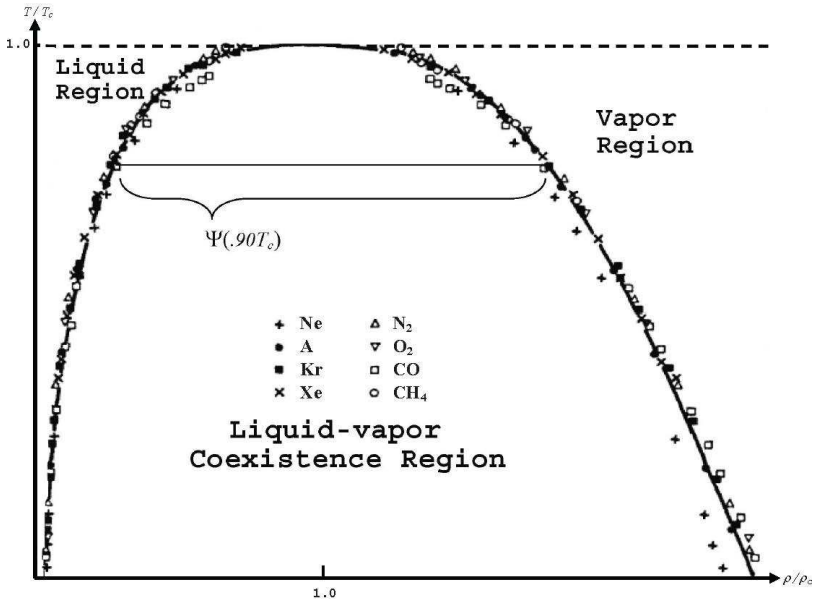


Figure 2.2: *Adapted plot of (Guggenheim, 1945) data rescaled for criticality.*

states at which various fluids transition from a liquid or vapor state to a "two phase" liquid-vapor coexistence region. Systems in states located in this latter region can be in liquid or vapor phases and (according to thermodynamics) maintains constant temperature as the density of the system changes. An important feature exhibited in figure 2.2 is that (after rescaling for the ρ_c and T_c of the respective molecules) the transition points of the each of the distinct substances near criticality appears to be well fit by a *single curve* referred to as the *coexistence curve*. This similarity in the coexistence curves best fitting diverse molecular substances can be characterized by a particular value β referred to as the *critical exponent* found in the following relation:

$$\Psi(T) \propto \left| \frac{T - T_c}{T_c} \right|^{\beta} \tag{1}$$

where the parameter $\Psi(T)$, called the *order parameter* tells us the width of the coexistence curve at a particular temperature value T. As depicted in figure 2.2, as T gets closer and closer to the critical temperature T_c from below, this width drops down eventually

vanishing at criticality. We can think of the critical exponent β as telling us about how rapidly such a vanishing occurs. As confirmed by the above data, this number turns out to be similar (in the neighborhood of $\beta \simeq .33$) for vastly different fluid substances.[1]

What is fascinating about examples such as this is not the universal (or "nearly" universal) regularity in physical systems. That uniform reliable regularities (viz. "universal laws") can be found to apply to numerous physical systems (though remarkable) is nothing new. The interesting part is that such uniform reliable behavior occurs *despite the fact that at least at one level of description the systems are so incredibly dissimilar*. From a level of description thought to be perhaps more "fundamental" than the gross state variables (P, T, and ρ) used to characterize thermodynamic systems, the various substances exhibiting similar critical exponent values have quite diverse descriptions: At the quantum mechanical level, for instance, the state vectors or density matrices representing the respective quantum mixtures will be incredibly distinct (e.g. close to orthogonal). Moreover, we need not go down to a quantum level of description to recognize the vast diversity. From a chemical perspective monotonic neon is different from a diatomic oxygen molecule, or an asymmetrical carbon monoxide molecule. We might hence expect surprise from a physicist or chemist since despite such vast differences in the ostensibly pertinent details at these levels of theorizing, the substances still share this observed similarity. This similarity despite such (speciously relevant) differences is what distinguishes the behavior across thermal systems as a kind of *universality phenomenon*. In the next section we begin a more explicit analysis of the concept's general application in physics.

Though the usage of the term originated in the study of thermal systems, universality has now been identified in a multitude of other domains. Over the past decade, Robert Batterman has argued in the philosophical literature that "while most discussions of universality and its explanation take place in the context of thermodynamics and statistical mechanics,... universal behavior is really ubiquitous in science" (Batterman, 2002). A (far from comprehensive) list of vindicating examples includes the clustering behavior found in contexts including non-thermal criticality patterns exhibited in avalanche and earthquake

[1]This similarity in the value of the critical exponent exists not only for thermal fluid systems, but also in describing the behavior of ferromagnetic systems in the neighborhood of a thermal state that can be analogously characterized as the critical point.

modeling (Kadanoff et al., 1989; Lise and Paczuski, 2001), extinction modeling in population genetics (Sole and Manrubia, 1996), and belief propagation modeling in multi-agent networks (Glinton et al., 2010). Batterman has discussed many examples of universality phenomena distinct from criticality phenomena, including patterns in rainbow formation, semi-classical approximation, and drop breaking(Batterman, 2002, 2005). Numerous non-criticality examples of universality have also been discovered in contexts such as the study of chaotic systems exhibiting "universal ratios" in period doubling (Feigenbaum, 1978; Hu and Mao, 1982), or the clustering similarities in models of cold dark matter halos found in astronomical observations (Navarro et al., 2004), to name a couple. In the next section we offer an explicit analysis of the concept's general application in physics.

2.2 The Same but Different: Analyzing Universality

The term *universality* is used in physics to describe cases in which broad similarities are exhibited by classes of physical systems despite possibly significant variations according to apparently "more fundamental" representations of the systems. Kadanoff (2000, p225) describes the term most generally as applying to those patterns in which "[m]any physically different systems show the same behavior." Berry (1987) has characterized it as the "way in which physicists denote identical behavior in different systems." Batterman (2002, p4) explains that the "essence of universality" can be found when "many systems exhibit similar or identical behavior despite the fact that they are, at base, physically quite distinct." Characterizations such as these reveal that the concept hinges on the satisfaction of the two seemingly competing conditions of displaying a particular *similarity* despite other (evidently irrelevant) *differences* in the systems at some level of description. To make this conceptual dependency explicit, we propose the following analysis of *universality phenomena*.

(UP): A class $X_{\mathcal{T}}$ of models of physical systems in a theoretical context $\mathcal{T}$ will be said to exhibit a *universality phenomenon* whenever the class can simultaneously meet the following two conditions:

(Sim) There exists a robust similarity in some observable behavior across

the physical systems modeled by members of $X_{\mathcal{T}}$.

(Var) This similarity in the behavior of members modeled in $X_{\mathcal{T}}$ is stable under robust variations of their state descriptions according to context $\mathcal{T}$.

The first thing to specify is what counts as a "class of models of physical systems in a theoretical context." In order to avoid complications associated with multiple (possibly not entirely equivalent) formulations of a full physical theory, **(UP)** is best analyzed in terms of the more restrictive notion of a theoretical context $\mathcal{T}$ which identifies within a given theory a particular formulation and variety of studied phenomena. Examples of different theoretical contexts in classical mechanics include the Hamiltonian versus the Lagrangian formulations, or in quantum mechanics we might distinguish between wave mechanics and operator mechanics.[2] A theoretical context may also restrict the phenomena considered by the total theory. For example, *source free* classical electrodynamics might be considered a distinct theoretical context within the full theory of classical electrodynamics which also models the effects of sources. In some cases it is possible for a theoretical context $\mathcal{T}$ to specify an entire theory uniquely, in other cases, a specification in terms of (potentially nonequivalent) formulations and specific phenomena types may be appropriate.

Given a particular theoretical context $\mathcal{T}$ of a universality phenomena, the expert will typically be able to identify pertinent state descriptions "according to context $\mathcal{T}$." For example, in classical electromagnetism the relevant state description may come in the form of fields specifying the flow of the source charges and the electromagnetic field values throughout a spacetime; in general relativity the metric and energy-momentum tensors might play this role; in thermodynamics, state descriptions may be parametrized by P, T, and ρ (or perhaps V and N), whereas in quantum statistical mechanics one may use density operators.

Satisfaction of (Sim) is primarily an empirical question. In order to claim that something universality-like is occurring, there must be an evident similarity in the class of systems exhibiting the phenomenon. This evident similarity need not be (directly) in terms

[2]Note, in both dichotomies there exist occasional circumstances or conditions such that the respective formulations can cease to be equivalent.

of any of the state descriptions used to characterize elements of $X_{\mathcal{T}}$. So for the paradigm example of the universality of phase transitions, (Sim) is satisfied once physicists recover sufficient empirical data of the kind depicted in figure 2.2. The robust similarity of (Sim) can be quantified in terms of the remarkable closeness of the critical exponents of these various systems even though the critical exponent parameter β may not necessarily be put in terms of the state quantities of $\mathcal{T}$ (e.g. chemistry or statistical mechanics).

Satisfaction of (Var) depends primarily on the size and most importantly the diversity of the models in class $X_{\mathcal{T}}$. The larger and more varied the members of class $X_{\mathcal{T}}$ with respect to the relevant state descriptions of $\mathcal{T}$, the more "stable under variations." If $X_{\mathcal{T}}$ is suitably rich with diverse members, then a member $x \in X_{\mathcal{T}}$ may be "mapped" to a rich variety of other members of $X_{\mathcal{T}}$ while still maintaining the very similarity shared by all members of $X_{\mathcal{T}}$ that allowed the class to satisfy (Sim). In the paradigm example of thermal universality, (Var) is satisfied by the fact that at the chemical or the statistical mechanics levels of description, the members in our class sharing this similar critical behavior are so diverse.

We note that the central concepts of *robust variation* and *robust similarity* on which (Var) and (Sim) respectively depend are not binary. Some universality phenomena may be "more robust" than other instances, in terms of both the "degree" of similarity displayed and the "degree" of variations that the systems can withstand while still exhibiting such similar behavior. The greater the robustness of the pertinent similarity in behavior across the class of systems and the more ($\mathcal{T}$-state) variation in the class, the more robust the universality is.[3] This non-binary dependence means universality may be subject to vagueness challenges in some cases. While certain examples, such as thermal criticality behavior and, as we argue, the clustering behavior of free-fall massive bodies around geodesic paths may be identified as determinant cases of universality, penumbral cases where it is unclear whether a candidate universality class is sufficiently similar and robust under variations may exist.

[3] Often this can be rigorously assessed by an appropriately natural norm, metric, topology, etc. defined on the state descriptions of $\mathcal{T}$. E.g. we might use some integration norm to quantify the difference between two (scalar) fields found in $X_{\mathcal{T}}$. The choice of appropriate norm, topology, etc. identifying differences in the members of $X_{\mathcal{T}}$ is directly dependent on the context $\mathcal{T}$.

3 The Geodesic Universality Thesis

In this section we reconsider the case of near-geodesic clustering observed in nature in terms of the **(UP)** analysis. In 3.1 we examine why such clustering qualifies as an example of a universality phenomenon. In 3.2 we then identify how the limit operation result of Ehlers and Geroch offers what we identify as a *universality explanation* of this clustering.

3.1 The Similarity and Diversity of Geodesic Universality

Consider a sequence of classes $(X^{\epsilon}_{\mathcal{GR}})_{\epsilon\in(0,s)}$ indexed by some sufficiently small error parameter $\epsilon \in (0, s)$. For fixed ϵ, the class $X^{\epsilon}_{\mathcal{GR}}$ consists of (local) solutions to Einstein's field equations:

$$T_{ab} = G_{ab} \tag{2}$$

where the energy-momentum field T_{ab} describes the flow of matter-energy and G_{ab} describes the "Einstein curvature" determined by the metric field g_{ab}. Moreover, each member of $X^{\epsilon}_{\mathcal{GR}}$ models some massive body whose spacetime path comes close to following a (timelike) curve γ that is close to actually being a geodesic (where these two senses of closeness are parametrized by respective functions monotonically dependent on the smallness of ϵ). With the **(UP)** analysis in hand, for a given degree of "ϵ-closeness" we can now ask if such a class $X^{\epsilon}_{\mathcal{GR}}$ satisfies the (Sim) and (Var) conditions in the context of general relativity theory purged of the canonical commitment to geodesic dynamics argued against in (Tamir, 2012).

The satisfaction of (Sim) is an empirical matter apparently well confirmed by centuries of astronomical data recovered from cases in which a relatively small body (a planet, moon, satellite, comet, or even a star) travels under the influence of a much stronger gravitational source. Examples involving non-negligible relativistic effects (like the Mercury confirmation) are of particular importance, but even terrestrial cases including Galileo and leaning towers or other (nearly) free-fall examples in determinately Newtonian regimes can count as confirming instances for certain ϵ-closeness values. Since observational precision is in-

evitably bounded, it is often claimed that the satellite, moon, planet, etc. indeed "follows a geodesic," despite the results of (Tamir, 2012). In such instances, the body is actually observed to come "close enough" to following a geodesic to warrant such equivocation. These instances hence confirm membership in a class $X^{\epsilon}_{\mathcal{GR}}$ for some ϵ threshold below the level of experimental precision or attention.

In order to appreciate the satisfaction of (Var), we must consider the relevant theoretical context of general relativity theory. State descriptions of physical systems according to the theory come in the form of the tensor fields T_{ab} and g_{ab}, related by the equations (2). Assuming we only consider (local) solutions to Einstein's equations, there exist six independent field components describing g_{ab} and so the matter-energy flow T_{ab}. In other words, from a fundamentals of relativity theory perspective, there are six physical degrees of freedom to how these bodies are described at each spacetime point.

Given the wealth of evident confirming instances falling under a class $X^{\epsilon}_{\mathcal{GR}}$ with suitable ϵ, there will be significant variation in terms of these degrees (even after rescaling) once we consider the significant differences in the density, shape and flow of the matter-energy of a planet, versus a satellite, asteroid, anvil, etc. In these "fundamental state description" terms, the diversity of the bodies in a given class $X^{\epsilon}_{\mathcal{GR}}$ will be quite significant. Despite this diversity, such bodies still satisfy the defining requirement of ϵ-closeness to following a geodesic. It is with respect to this diversity in these degrees of freedom (of the energy-momenta/gravitational influences of the "near-geodesic following bodies" of members in $X^{\epsilon}_{\mathcal{GR}}$) that a "robust stability under variations" can be established in accordance with (Var).

So, according to our **(UP)** analysis, such near-geodesic clustering observed in nature constitutes a *geodesic universality phenomenon*. However, meeting the conditions of the analysis depends entirely on the truth of the above made *empirical* claims about the existence of bodies well modeled by members of the respective $X^{\epsilon}_{\mathcal{GR}}$ classes for a suitable range of ϵ values, and that the bodies in each class are so fantastically diverse from the perspective of their T_{ab} (g_{ab}) fields. In the next section we turn to the more *theoretical* question of understanding how such geodesic universality is possible in general relativity, by considering the properties of the classes $(X^{\epsilon}_{\mathcal{GR}})_{\epsilon\in(0,s)}$ in terms of an important geodesic

result of Ehlers and Geroch (2004).

3.2 Explaining Geodesic Universality

We have now formulated the geodesic universality thesis in the context of general relativity as an empirically contingent claim about classes of the form $X^{\epsilon}_{\mathcal{GR}}$ whose members model a physical system such that the path of some body counts as ϵ-close to being geodetic without violating Einstein's field equations. We have also given a plausibility argument suggesting why observational data already obtained by experimentalists confirms this empirical hypothesis. Moreover, given such confirmation and the diversity of the energy-momenta of the respective bodies, membership in some $X^{\epsilon}_{\mathcal{GR}}$ will be sufficiently stable under significant variations of the fundamental state descriptions of the theory to satisfy (Var). A remaining theoretical question must now be answered: *How can the systems exhibiting this universality phenomenon behave so similarly while being so different at the level of theoretical description fundamental to general relativity?*

Geodesic universality can be explained by appealing to an important "limit proof" of the geodesic principle discussed in (Tamir, 2012). It was argued there that Ehlers and Geroch (2004) are able to deduce the "approximate geodesic motion" of gravitating bodies with relatively small volume and gravitational influence, by considering sequences of energy-momentum tensor fields with positive mass of the form $(\underset{(i,j)}{T}{}_{ab})_{i,j\in\mathbb{N}}$, referred to as "EG-particles." The spatial extent and gravitational influence of these EG-particles can be made arbitrarily small by picking sufficiently large i and j values respectively. The theorem of (Ehlers and Geroch, 2004) entails that if for a given curve γ there exists such an EG-particle sequence, then by picking a large enough j, γ comes arbitrarily close to becoming a geodesic in a spacetime containing the $\underset{(i,j)}{T}{}_{ab}$ instantiated matter-energy.

Specifically, let $(\underset{(i,j)}{g}{}_{ab})_{i,j\in\mathbb{N}}$ be the sequence of metrics that couple to these $(\underset{(i,j)}{T}{}_{ab})_{i,j\in\mathbb{N}}$ according to (2) in arbitrarily small neighborhoods $(\mathcal{K}_i)_{i\in\mathbb{N}}$ of γ, containing the support of the respective $\underset{(i,j)}{T}{}_{ab}$. Then if for each i, as $j\to\infty$ the $\underset{(i,j)}{g}{}_{ab}$ approach a "limit metric" g_{ab} in the $\mathscr{C}^1(\mathcal{K}_i)$ topology, which keeps track of differences in the metrics and their unique connections, then the curve γ approaches geodicity as $j\to\infty$.

To understand the impact of the theorem for our universality classes $(X^{\epsilon}_{\mathcal{GR}})_{\epsilon\in(0,s)}$, we need to appreciate the kind of limiting behavior established by Ehlers-Geroch. The limit result essentially establishes a kind of "ϵ-δ relationship" between, **(a)** how "nearly-geodetic" we want the curve γ to be, and **(b)** how much we need to bound the gravitational effects of the body on the background spacetime.[4] That is to say, the Ehlers-Geroch limit result can be thought of as telling us that "for every degree of *ϵ-closeness* to geodicity we want the bodies' path to be, there exists a *δ-bound* on the gravitational effect of the body that will keep the path at least that close to geodicity." The important thing to observe about this ϵ-δ interplay is that though the limiting relationship *does* require imposing a δ-bound on the perturbative effects of the body, it does not impose any *specific constraints on the details* of how the matter-energy of the body flows within the ϵ-close spatial neighborhood of the curve, nor how the metric it couples to specifically behaves. So though the metric is "bounded" within a certain δ-neighborhood of the limit metric, the particular details of the tensor values, the corresponding connection, and especially the curvature have considerable room for variation so long as they stay "bounded in that neighborhood."

This relationship established by the Ehlers-Geroch theorem hence gives us a kind of *details-free* way of understanding the diverse populations of our respective universality classes $(X^{\epsilon}_{\mathcal{GR}})_{\epsilon\in(0,s)}$. In effect the Ehlers-Geroch limiting relationship highlights that for each $X^{\epsilon}_{\mathcal{GR}}$ class, there exists a particular δ-bound around a limit metric with some geodesic anchor γ such that any body coupling to a metric that stays within that bound (in addition to remaining spatially close enough to γ) satisfies the relevant ϵ-closeness part of the requirements for membership in $X^{\epsilon}_{\mathcal{GR}}$. But as we just emphasized, *falling under this δ-bound does not impose specific constraints on the detailed values of the energy-momenta or metric fields*. In other words, membership in the universality class $X^{\epsilon}_{\mathcal{GR}}$ is possible as long as the body is a massive solution to Einstein's equations, and its gravitational effect and extent are sufficiently bounded in the right way, but beyond these requirements the specific details concerning "what the gravitational effect does below those bounds" are

[4]For purposes of exposition, we characterize the established relationship as an "ϵ-δ relationship," suggesting that the closeness relations in question have been quantified, the actual Ehlers-Geroch result is formulated (primarily) in topological terms. See (Gralla and Wald, 2008, §3-5) for a more explicitly quantified approach.

irrelevant. Hence, the limit behavior established by the Ehlers-Geroch theorem explains how the ϵ-clustering near geodesic anchors is possible despite significant differences in the energy-momenta of our near-geodesic following bodies: So long as the bodies' gravitational influences are bounded in the right way their (positive) matter-energy can vary as much as we like under those bounds.

4 Explanation without Reification

Before concluding there remains a potential challenge concerning how we can endorse any kind of geodesic "idealization" thesis if the actual geodesic motion of massive bodies is incompatible with Einstein's theory. Recall, while explaining how the classes $X^{\epsilon}_{\mathcal{GR}}$ whose respective members are "ϵ-close" to geodesic following models could be so diverse, we needed to take the "geodesic limit" of the metrics $(\underset{(i,j)}{g}{}_{ab})_{i,j\in\mathbb{N}}$ coupling to the EG-particles $(\underset{(i,j)}{T}{}_{ab})_{i,j\in\mathbb{N}}$ in accordance with the equations (2).[5] By taking such a "geodesic limit" to identify the diversity of our $X^{\epsilon}_{\mathcal{GR}}$ classes, haven't we made an "essential" appeal to the kind of pathological models precluded by Einstein's field equations?

The answer to this challenge is that though appreciating the kind of ϵ-δ interplay in the appropriate neighborhoods of the geodesic limit was essential to our explanation of geodesic universality, the role played by the limiting *geodesic anchor model* does not require us to reify the idealization or make it representative of any physical system in Einstein's theory. Even though there are significant complications associated with what happens *at* the geodesic limit **(1)** the ϵ-δ behavior of the systems has a well-defined mathematical structure (the $\mathscr{C}^1$ topologies defined for each spacetime neighborhood of γ) describing the *approach* to the limiting anchor model, and **(2)** the behavior of the models in $X^{\epsilon}_{\mathcal{GR}}$, which are "close but not identical to" a geodesic anchor model, still obey Einstein's theory. A geodesic anchor model establishes (as the name suggests) a kind of *anchor* for the (topological) neighborhoods within which the elements of the respective

[5]Note, though the $({}_{(i,j)}g_{ab})_{i,j\in\mathbb{N}}$ converge to a well defined "geodesic limit" (in the $\mathscr{C}^1$ topologies) the coupled energy-momentum tensors $({}_{(i,j)}T_{ab})_{i,j\in\mathbb{N}}$ may not. Moreover, even if they do converge in a physically salient and independently well-defined way, at the limit they must either fail to obey (2) or vanish. For a detailed discussion see (Tamir, 2012, §4).

$X^{\epsilon}_{\mathcal{GR}}$ can be said to cluster. However, using these models as anchors to identify the points around which the *actual solutions* to Einstein's equations cluster does not require that the anchors themselves be admitted in $X^{\epsilon}_{\mathcal{GR}}$.

In contrast to more traditional "idealizations," universality phenomena are about the *group behavior of classes of* $X_{\mathcal{T}}$ not *individual systems*. For non-universality idealizations severe pathologies can be detrimental because they render *the sole idealized model* theoretically inapposite. With universality, however, the existence of a pathologically idealized model "close to but excluded from" a universality class need not entail that members of the class are likewise poorly behaved. Moreover, if a topological clustering "near to" an idealized model has physical significance (as with the $\mathscr{C}^1$ topologies), such proximity may allow inferences about the well-behaved classes without molesting their admissibility according to the laws of $\mathcal{T}$.

This is precisely what occurs with geodesic universality. Members of a class $X^{\epsilon}_{\mathcal{GR}}$ can take advantage of their closeness to the geodesic anchor models without "contracting" the pathologies occurring *at* the actual geodesic limits. Moreover, we were able to *explain* such ϵ-closeness by appealing to what we characterized as the "specific details irrelevant" δ-closeness in the $\mathscr{C}^1$ topologies. Since we are talking about geodesic universality, we are able to infer directly from such ϵ-closeness that the relevant bodies modeled by the members of $X^{\epsilon}_{\mathcal{GR}}$ are *close* to following a geodesic in the relevant physical senses defined when we constructed the classes.

5 Conclusion

While the incompatibility result of (Tamir, 2012) entails that the geodesic principle strictly interpreted must be rejected at the fundamental level, in this paper we have argued that reinterpreting the role of geodesic dynamics as a universality thesis is both viable and coherent with contemporary general relativity. By developing an analysis of universality phenomena in physics, we saw that the widespread geodesic clustering of a rich variety of gravitating, free-fall, massive bodies actually observed in nature qualifies as a geodesic universality phenomenon.

Not only can this approximation of geodesic dynamics be recovered in the form of such a geodesic universality thesis, but by reconsidering the implications of limit operation proofs of the principle, we were able to generate a universality explanation for why we can expect such a remarkable clustering of these gravitating bodies despite the fact that from the perspective of their more fundamental relativistic descriptions (the energy-momentum field and its gravitational influence) they may be incredibly dissimilar. We concluded with a defense of our appealing to pathological geodesic anchor models in explaining the universality clustering. Unlike more traditional forms of approximation or idealization, as revealed by the **(UP)** analysis, when it comes to universality phenomena, the claim is about *the group behavior* of entire classes of models, not individual idealizations. Hence, in the case of universality, it is possible to take advantage of relevant types of mathematical proximity to pathological anchors without actually infecting the members of the class with the illicit behavior. Moreover, when the right kind of (topological) closeness is employed it may be possible to draw inferences and gain knowledge about the physical properties of modeled systems thanks to this proximity of their models to the pathological anchor.

References

Batterman, R., 2002. The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence. Oxford University Press, USA.

Batterman, R., 2005. Critical Phenomena and Breaking Drops: Infinite Idealizations in Physics. Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics 36 (2), 225–244.

Berry, M., 1987. The Bakerian Lecture, 1987: Quantum Chaology. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences 413 (1844), 183.

Ehlers, J., Geroch, R., 2004. Equation of Motion of Small Bodies in Relativity. Annals of Physics 309 (1), 232–236.

Feigenbaum, M., 1978. Quantitative Universality for a Class of Nonlinear Transformations. Journal of Statistical Physics 19 (1), 25–52.

Glinton, R., Paruchuri, P., Scerri, P., Sycara, K., 2010. Self-Organized Criticality of Belief Propagation in Large Heterogeneous Teams. Dynamics of Information Systems 40, 165–182.

Gralla, S., Wald, R., 2008. A Rigorous Derivation of Gravitational Self-force. Classical and Quantum Gravity 25, 205009.

Guggenheim, E., 1945. The Principle of Corresponding States. The Journal of Chemical Physics 13, 253.

Hu, B., Mao, J., 1982. Period Doubling: Universality and Critical-point Order. Physical Review A 25 (6), 3259.

Kadanoff, L., 2000. Statistical Physics: Statics, Dynamics and Renormalization. World Scientific Publishing Co.

Kadanoff, L., Nagel, S., Wu, L., Zhou, S., 1989. Scaling and Universality in Avalanches. Physical Review A 39 (12), 6524–6537.

Lise, S., Paczuski, M., 2001. Self-organized Criticality and Universality in a Nonconservative Earthquake Model. Physical Review E 63 (3), 036111.

Navarro, J., Hayashi, E., Power, C., Jenkins, A., Frenk, C., White, S., Springel, V., Stadel, J., Quinn, T., 2004. The Inner Structure of Λ CDM Haloes–III. Universality and Asymptotic Slopes. Monthly Notices of the Royal Astronomical Society 349 (3), 1039–1051.

Sole, R., Manrubia, S., 1996. Extinction and Self-organized Criticality in a Model of Large-scale Evolution. Physical Review E 54 (1), 42–45.

Tamir, M., 2012. Proving the principle: Taking geodesic dynamics too seriously in Einstein's theory. Studies in History and Philosophy of Modern Physics 43, 137–154.

New Work for a Theory of Emergence

Abstract

Many discussions of emergence focus on the putatively irreducible causal powers of emergents. I argue that the question of whether emergents are irreducible causes should be postponed in favor of the question of whether emergents are natural. David Lewis contends that scientifically interesting properties are natural: they account for resemblances and distinguish causally relevant and irrelevant properties. I consider bimanual coordination, as described by the HKB model, as an example of natural emergence. I conclude that emergent properties are those that systemic processes have when they exhibit "instability" and argue that instability is the natural property that unites these processes.

1. Introduction.

Complexity scientists use "emergence" to describe dynamical processes in open systems that produce spontaneous reorganization of their components. Philosophical discussions tend to focus on the putatively irreducible causal powers of emergents relative to the causal powers of their subvenient micro-constituents. I argue that the question of whether they are irreducible causes should be postponed in favor of the question of whether emergents are natural. Lewis (1983) argues that scientifically interesting properties are natural: they account for resemblances among things and distinguish the causally relevant and irrelevant properties. If emergent properties are scientifically interesting, then they are natural. I consider bimanual coordination (described by the HKB model) as an example of emergence

in a natural system. Inspired by Rueger (2000), I define emergence as a property that systemic processes have when they begin to exhibit "instability", and argue that instability is the natural property that unites these processes.

2. Emergents as Irreducible Causes.

Kim thinks that an ontological interpretation of emergence requires that emergent properties have either synchronic or diachronic "downward causal influence". This interpretation requires that

> the emergents bring into the world new causal powers of their own, and, in particular that they have powers to influence and control the direction of the lower-level processes from which they emerge. (1999, 5-6)

Kim finds the notion of synchronic downward causal influence unintelligible because it violates the principle that if an object is caused to have a causal power at a time *t*, then it cannot exercise that causal power at *t* (1999, 147). Exercising a causal power happens later than the time of acquisition.

Diachronic downward causal influence, according to Kim, creates an intolerable causal competition between emergents and their non-emergent bases. Kim queries "if an emergent, M, emerges from basal condition P, why can't P displace M as a cause of any putative effect of M?" (1999, 147) Reductionists argue that an emergent property cannot causally compete with its subvening micro-based physical property. There are several replies that attempt to salvage the causal efficacy of emergents without positing entirely distinct

properties. For example, a Pereboomian account of emergence might suggest that token emergent causal powers are materially constituted by token microphysical-based causal powers (2011, 141). Shoemaker's (2001) account of mental causation suggests that emergent properties are determinables, the causal powers of which are a subset of the causal powers of their microphysical realizers. I think this debate is a bit premature. If we want an ontological interpretation of emergence on which emergents cause anything at all, then we need some reason to think that they are not grueish and gerrymandered. Figuring out the precise relation that emergent properties have to micro-based properties should be postponed until we have some account of that. Whatever is included in the category of "emergent", the matter of determining whether or not the emergents are gerrymandered is to be settled by paying attention to systematic empirically theorizing.

3. What Would It Mean for Emergence to be Natural?

Lewis (1983, 1986) argues that accounting for the kinds of similarity that distinguishes causally relevant and irrelevant properties requires a metaphysical theory that either includes universals as real primitives or provides an adequate nominalism that does the work that universal would do. Both universals and nominals, according to Lewis, would perform this work by permitting a distinction between natural and unnatural properties:

> Because properties are so abundant, they are undiscriminating. Any two things share infinitely many properties, and fail to share infinitely many others. That is so whether the two things are perfect duplicates or utterly dissimilar. Thus properties do nothing

> to capture facts of resemblance.... Likewise, properties do nothing to capture the causal powers of things. Almost all properties are causally irrelevant, and there is nothing to make the relevant ones stand out from the crowd. (1983, 346)

The perfectly natural properties, for Lewis, would be coupled so tightly to resemblance that two things have exactly the same natural properties just in case they are qualitative duplicates (356). Imperfectly natural properties come in varying degrees because they are built up from suitably "close-knit" perfectly natural properties (ibid., 347). To account for the nature of this close-knit relationship Lewis suggests that perfectly natural properties could stand in resemblance relations to each other. By doing so they would constitute families of imperfectly natural properties, which may themselves constitute families of even less perfectly natural properties. Fundamental sciences study the perfectly natural properties and special sciences concentrate on families of imperfectly natural properties. All of the disjunctive properties are unnatural and scientifically uninteresting because they do not comprise resembling properties.

It is not entirely clear how closely imperfect natural properties must resemble in order to qualify as the domain of a special science. However, I think Lewis's suggestion here is even more provocative than it appears on the surface. A property that falls under the domain of a particular science by being a member of a resemblance class might also be cross-listed under a different science if it falls under a different resemblance class. Such properties may compose the domain of an interdisciplinary science. Suppose that phenomena across a wide range of traditional fields from chemistry to economics could be accurately simulated using

the same mathematical model and that the predictive consequences of this model were all well confirmed. Given that there is an interesting association between duplicability and simulability, this would suggest that there is enough resemblance between these phenomena to compose a domain that deserves its own systematic investigation – especially if there are a sparse number of law-like generalizations that govern a wide range of these phenomena. Complexity Science is an interdisciplinary science of just this sort, and emergence is a central part of its purview. As Bedau remarks, "the models in complexity science are typically described as emergent, so much so that one could fairly call the whole enterprise the science of emergence" (2002, 5).

Lewis distinguishes genuine events from spurious events in order distinguish the causally relevant and irrelevant events (1983, 369). Genuine events have natural properties, which makes them fit to figure into laws and causal relations. Spurious events, on the other hand, are unfit for this task. Like *being grue*, the properties of spurious events are gerrymandered. The natural/unnatural distinction explains what makes the causally relevant properties stand out from the rest. Lewis also stresses that the coupling of causally relevant events to natural properties explains why "scientific investigation of laws and of natural properties is a package deal…so that laws and natural properties get discovered together" (ibid., 368). He takes it as uncontroversial that causation involves laws.[1] Whether or not they require laws, causal explanations cite natural properties. Indeed, the notion of invariance is

[1] As Woodward (2003) argues, it is more likely that causes are defined by change-relating invariances, which may come in degrees of invariance.

wrapped up with the notion of resemblance. If an invariant generalization remains invariant over a fairly wide range of different initial conditions and different types of intervention, then it is reasonable to think that it generalizes over the behavior of natural properties because those are the properties that account for resemblance between like cause and like effect. If dynamical models of complex causal systems have explanatory power then the properties they cite are natural properties. As I argue below, the HKB model explains a variety of surprising properties of coordinated movement.

4. Emergence, Resemblance, and Causal Explanation.

During the last sixty years Complexity Science blossomed out of research originally grounded in General Systems Theory. Because complexity scientists are interested in diachronic processes that display emergence, they look for the similarities relevant to emergence in the dynamics of these processes. Is there a suitably related family of properties that corresponds to these different dynamical processes? I argue that there is such a family. To justify such a sweeping claim is a daunting task, but I can give partial justifications by detailing a dynamical model that illuminates emergence in nature and by describing how other such models do similar work.

The HKB model is a first-order differential equation introduced by Kelso, Hakken, and Bunz to model the behavior of the coupled oscillator systems involved in bimanual coordination (Kelso 1995). Here is a simple experiment that shows how the oscillator governing rhymic motions in your right and left hands exhibit emergent properties when

coupled together. Snap with the fingers on your right hand in a rhythmic fashion and then snap with your left hand at the same time. Your snapping fingers will be oscillating in-phase. If you change the snapping rhythm to anti-phase, then you will be snapping a regular beat with a fingersnap on one hand followed by a fingersnap on the other. Only in-phase and anti-phase rhythms are stable. Because the system has two stable states, the system is *bistable*. Both the in-phase oscillations and the anti-phase oscillations are stable up to a certain frequency. If the frequency crosses a certain critical threshold, the anti-phase pattern becomes unstable. If you keep increasing the rate, you will find yourself snapping in-phase. This is a kind of *phase transition* known as a *bifurcation*. The in-phase state is globally stable. As fast as you can snap, you can snap in-phase. Fluctuations in the systems may briefly pull you away from the in-phase pattern, but your snapping behavior will be attracted to the in-phase pattern and repelled away from all others. The *relative phase* of the oscillations is the variable that captures the stable and unstable states of coordination. This simple discovery is the basis for the HKB model of coordination.

Eq.1 gives an idealized version of the HKB equation that ignores the intrinsic tuning of oscillators and systemic noise.

Eq.1 $\varphi^* = -a \sin(\varphi) - 2b \sin(2\varphi)$

The parameter φ indexes the relative phase of the coupled oscillator system, and φ^* is the change in relative phase over time. In Eq.1 parameters a and b are the coupling strength coefficients and $-a \sin(\varphi) - 2b \sin(2\varphi)$ is the coupling term. Decreasing the ratio of b/a is equivalent to shortening the period of rhythmic coordination. Because the oscillators are

time-symmetric and cyclical, we can represent their periods as a 360° cycle where possible relative phase φ ranges from –π to π.

The below plots (taken from Kelso 1995, 55) depict the potential function of the HKB model.

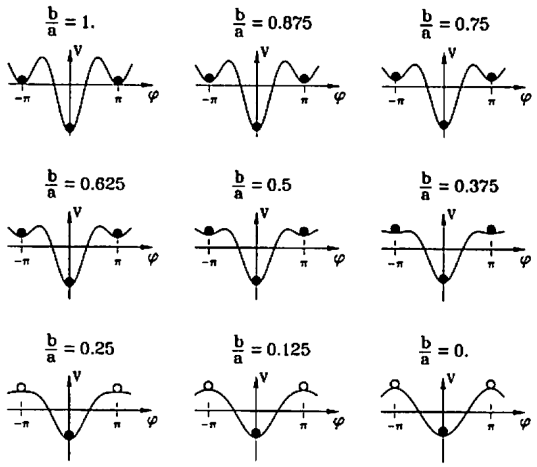


Figure 1
(taken from Kelso 1995, p. 55)

Solid black balls indicate stable attractors and open white balls indicate unstable repellers. At $b/a > .25$ the system has three stable states corresponding to the two kinds of anti-phase patterns and the in-phase pattern. At a critical value of $b/a = .25$ there is a phase transition to having only one stable state, which is the in-phase pattern. Why would a gradual increase in frequency of snapping lead to a sudden change in the kinds of stable patterns that one is able to snap? The key to the model is the parameter that measures relative phase. It is what Kelso calls a collective variable, or order parameter, because it measures the cooperative macro-level behavior of the system. Kelso insists that "to understand coordinated behaviors as self-

organized, new quantities have to be introduced beyond the ones typical of the individual components." (1995, 52) When the snapping oscillators are coupled, they constitute a new cooperative system. No account of the behavior can be given at the level of single snapping fingers. You need to introduce a new adjustable parameter in order to understand how and why the coupled system behaves as it does. The model also makes two novel predictions of expected phenomena, *critical fluctuations* and *critical slowing down*, not previously recognized, but since confirmed (ibid., 71).

The most interesting feature of the HKB model is the breadth of phenomena that can be understood using the model. Finger snaps are an arbitrarily chosen example of the types of bimanual coordination that the equation models. Among the other factors that have been experimentally demonstrated to break the symmetry of HKB are handedness/ hemispheric asymmetry (Treffner and Peter 2002), mutilimb coordination (Mechsner et. al. 2001), attentional allocation, and speech-hand coordination (Treffner and Peter 2002).

We can now return to the questions that motivated this discussion:

1) Is there a suitably related family of properties that corresponds to self-organizing complex systems such that they are grouped under a natural property?

2) Is there reason to think that such properties figure into causes and can be cited in causal explanations?

3) If so, does this natural property ground the distinction between systemic process that are ones of emergence and ones that are not?

I suggest an affirmative answer to the first two questions. The HKB model's broad applicability implies that it describes systemic processes that share united features. This is why the model applies to a wide variety of psychological and biomechanical phenomena in humans and non-human animals. Other complex oscillatory models, such as the van der Pol model that accounts for limit cycle behavior in both neural polarization and electrical circuits employing vacuum tubes, are similarly interesting because of their interdisciplinary applications. Likewise, these models give us the kinds of change-relating invariant generalizations that indicate systemic processes with causal powers. In particular, the need to introduce order parameters to describe them signals that there is a unique feature of these kinds of self-organizing complex systems that unites them into a kind that has scientific significance. My answer to the third question is affirmative as well, but to properly address it I first need to do some conceptual housekeeping.

5. Conclusion.

Systems that conform to the HKB model resemble each other in a way that is appropriate to being grouped under a natural kind. But the HKB conforming systems are only one type of system that exhibits non-liner phase transitions. Some (e.g. limit cycles) go from having two stable states to having one stable state. Others (e.g. Rayleigh-Benard convection) go from one stable state to two or more possible stable states. These kinds of systems tend to teeter on the edge of chaos until something pushes them over into complete disorder; they achieve their greatest level of complexity just before they lose structure altogether. Phase transitions,

I suggest, characterize only one kind of dynamical process united under the property of emergence. What then characterizes emergence and determines whether or not *being a process of emergence* counts as a natural property? For all that I have argued so far, emergence might characterize a gerrymandered set of properties that are themselves natural, but not constitutive of a natural family of properties.

Various definitions and desiderata for emergence have been offered in both the scientific and philosophical literature (Wimsatt 2007, Humphreys 2008, Bedau 1997 and 2002, Holland 1998, Ryan 2007). Reuger (2002) intimates a notion of emergence that fits well with the observations made in this essay.

> Suppose that the system's dynamics is characterized by an equation of motion with a control parameter p. If the phase space portrait stays qualitatively the same under perturbations of the dynamics, i.e., small variations in the value of p, the system is structurally stable. If the perturbation generates a qualitatively different portrait of trajectories, the system is structurally unstable. (Rueger 2000, 472-473)

The property of being unstable, in this sense, should be defined as counterfactual supporting. A systemic process exhibits instability if it has an order parameter that would create a drastic change in response to certain slight variations in a control parameter. We see this in the bimanual coordination cases. For example, slight variations in the rate of snapping lead quickly to rapid changes in the kind of coordination exhibited. When a system acquires a collective level property that is only described by a novel order parameter, then it increases in complexity; this is so whether or not the phase transition ever actually occurs.

My conceptual housekeeping is now complete and I can offer a definition of emergence that reveals how emergence constitutes a category that unites various resembling processes under the property of *having instability*.

Def.1 *A systemic process S exhibits instability iff there are small changes in the value of S's control parameter(s) that would lead to drastic changes in the value of S's order parameter(s).*

Def.2 *A systemic process S is a process of emergence iff S at time t does not exhibit instability I and at some later time t* S begins to exhibit instability I.*

There is one thing to note about this definition, it does not imply that losing an instability is a process of emergence. When systems plunge into chaos that is not emergence because it constitutes a loss of complexity. Despite the terminology, this account entails that falling into chaos entails a decrease of instability. The process of increasing in instability is an interesting property. I suggest that it is a good candidate for being a natural property because it unites instances of emergence under a genuine resemblance relation. Given that emergence is a natural property of the evolution of complex systems, and that having a natural property entails having causal powers, it may now reasonably be asked whether or not those causal powers are reducible. Whatever the answer to that question may be, both an adequate realism and an adequate nominalism about emergence should take account of its status as a natural property.

References

Batterman, Robert. 2002. *Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford: Oxford University Press.

Bedau, Mark. 1997. "Weak Emergence." In *Philosophical Perspectives: Mind, Causation, and World*, vol. 11, ed. James Tomberlin, 375-399. Oxford: Oxford University Press.

_____. 2002. "Downward Causation and Autonomy in Weak Emergence." *Principia* 6: 5-50.

Chemero, Anthony. 2009. *Radical Embodied Cognitive Science*. Cambridge, Mass.: MIT Press.

Haken, Hermann. 1983. *Synergetics: An Introduction*. Berlin, Germany: Springer.

Holland, John. 1998. *Emergence: From Chaos to Order*. Oxford: Oxford University Press.

Humphreys, Paul. 2008. "How Properties Emerge." In *Emergence: Contemporary Reader in Philosophy of Science*, eds. M. Bedau and P. Humphreys, 111-126. Cambridage, Mass.: MIT Press.

_____. 2008. "Synchronic and Diachronic Emergence." *Minds and Machines* 18: 584-594.

Jaegwon, Kim. 2000. *Mind In a Physical World: Essays on the Mind-Body Problem and Mental Causation*. Cambridge, Mass.: MIT Press.

_____. 1999. "Making Sense of Emergence." *Philosophical Studies* 95: 3-36.

_____. 2006. "Emergence: Core Ideas and Issues." *Synthese* 151: 547-559.

_____. 2009. "Supervenient and Yet Not Deducible: Is There a Coherent Concept of Ontological Emergence?" In *Reduction: Between the Mind and the Brain*, eds. Alexander Hicke and Hannes Leitgeb. London: Ontos Verlag.

_____. 2010. "The Layered Model of the World." *Essays in the Metaphysics of Mind*. Oxford: Oxford University Press.

Kelso, J.A. Scott. 1995. *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, Mass.: MIT Press.

Lewis, David. 1983. "New Work For a Theory of Universals." *Australasian Journal of Philosophy* 61: 343-377.

_____. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Mechsner, Franz, David Kerzel, Günther Knoblich, and Wolfgang Prinz. 2001. "Perceptual Basis of Bimanual Coordination." *Nature* 413: 69-73.

Pereboom, Derk. 2001. *Consciousness and the Prospects of Physicalism*. Oxford: Oxford University Press.

Reuger, Alexander. 2001. "Robust Supervenience and Emergence." *Philosophy of Science* 67: 466-491.

_____. 2001. "Physical Emergence, Diachronic and Synchronic." *Synthese* 124: 297-322.

Ryan, Alex. 2007. "Emergence is Coupled to Scope, Not Level." *Complexity* 13: 67-77.

Shoemaker, Sidney. 2001. “Realization and Mental Causation.” In *Physicalism and its Discontents*, eds. Carl Gillett and Barry Loewer, 74-98. Cambridge: Cambridge University Press, .

Treffner, P. and M. Peter. 2003. “Intentional and Attentional Dynamics of Speech-hand Coordination.” *Human Movement Science* 22: 641-697.

Wimsatt, William. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press.

Woodward, James. 2003. *Making Things Happen: a Theory of Causal Explanation*. New York: Oxford University Press.

Models, Sherlock Holmes and the Emperor Claudius

Abstract

Recently, a number of authors have suggested that we understand scientific models in the same way as fictional characters, like Sherlock Holmes. The biggest challenge for this approach concerns the ontology of fictional characters. I consider two responses to this challenge, given by Roman Frigg, Ronald Giere and Peter Godfrey-Smith, and argue that neither is successful. I then suggest an alternative approach. While parallels with fiction are useful, I argue that models of real systems are more aptly compared to works that portray real people, like the Emperor Claudius. This approach will allow us to avoid problems with fictional characters.

Word count: 4468 words (and 2 figures equivalent to approx. 200 words)

1. Introduction

Modelling forms an important part of scientific practice. It also presents us with a number of ontological puzzles. Consider the standard Newtonian model of the orbit of the earth. This model makes many simplifying assumptions: it assumes that the sun and earth are perfect spheres, for example, and that they are isolated from the other planets in the solar system. These assumptions are known to be false of the sun and earth. Indeed, no actual, concrete objects satisfy these assumptions. And yet scientists often talk as if there were such objects and as if they can find out about their properties. A scientist might say that the model consists of two spheres with homogenous mass distribution, for example, or she might discover that the orbit of the earth in the model is perfectly elliptical.

Let us call the various assumptions and equations that scientists write down when they formulate a model the model description (Godfrey-Smith 2006; Weisberg 2007). When she uses the Newtonian model, the scientist wishes to understand a real system, namely the sun and earth. But not all models are like this. For example, a predator-prey model might invite us to consider a population consisting of two species, predator and prey, whose numbers are governed by certain equations, without claiming to represent any real population out in the world. And yet, even in these cases, scientists talk as if the model were an object whose behaviour they are investigating. For example, they might discover that in certain models general pesticides act to increase the proportion of prey to predator (Weisberg 2007, 223). Notice that often the very same model description is put to different uses. We might write down the equation for a simple harmonic oscillator simply in order to explore the properties of such a system, or we might use it to understand the motion of a pendulum or a chemical bond.

Modelling thus presents us with certain ontological puzzles. How are we to make sense of the fact that a large part of scientific practice involves talking and learning about things that do not exist? One way to answer these questions is to insist that, while no actual, concrete object satisfies the scientists' model description, there is some other object that does satisfy it. According to Ronald Giere (e.g. 1988), for example, theoretical models are abstract objects defined by scientists' modelling assumptions. While this view has seemed attractive to many, it is not without problems. For example, Martin Thomson-Jones (2010) asks how the abstract objects posited by Giere's account can possess the spatiotemporal properties we appear to attribute to models, such as following an elliptical orbit or oscillating with a certain time period (see also Hughes 1997; Godfrey-Smith 2006).

Recently, a number of authors have suggested that, rather than abstract objects, theoretical models should instead be understood in the same way as fictional characters, like Sherlock Holmes. The aim of this paper is to examine this proposal in detail. The most obvious challenge for such an approach concerns the longstanding controversy over the nature of fictional characters (Section 2.1). I shall consider two ways in which proponents of the view have sought to respond to this challenge, and argue that neither response is successful (Sections 2.2 and 2.3). I will then suggest an alternative approach. While parallels with fiction are useful, I will argue that models of real systems are more aptly compared to works that portray real people, like the Emperor Claudius (Section 3.1). This approach will allow us to avoid problems with fictional characters (Section 3.2).

2. The Indirect Fictions View

2.1. Models and Fiction

A number of authors have been struck by apparent parallels between the ontology of models and fiction (e.g. Godfrey-Smith 2006; Thomson-Jones 2007; Contessa 2010; Frigg 2010a, 2010b; on models and fiction in general, see Suárez 2009). Consider the following passage from The Hound of the Baskervilles:

> Holmes leaned forward in his excitement and his eyes had the hard, dry glitter which shot from them when he was keenly interested. (Conan Doyle 1902/2003, 22)

Like scientists' model descriptions, it seems, there is no actual, concrete object that this passage describes: there is no real, flesh-and-blood detective that satisfies the description Conan Doyle gives of Holmes. And yet, just as scientists talk as if there were objects that satisfied their model descriptions, so we talk as if there were a Sherlock Holmes: we say that Holmes is highly intelligent, that he smokes a pipe and plays the violin. We saw above that some have criticised Giere's view on the grounds that we often ascribe spatiotemporal properties to models. We certainly have no problem attributing spatiotemporal properties to fictional characters: we say that Holmes is tall and that he lived at 221B Baker Street.

These observations motivate what I will call the indirect fictions view of modelling (figure 1).[1] According to this view, scientists' model descriptions give rise to what are called

[1] Here I use 'indirect' in a different sense to Michael Weisberg (2007). Weisberg uses the term 'indirect' to describe the activity of modelling, in order to distinguish it from other forms of theorising. I use 'indirect' and 'direct' to distinguish between two different

model systems (or sometimes simply models), and these model systems are to be understood in the same way as fictional characters like Sherlock Holmes. When scientists represent a real system they do so by establishing some form of representation relation between the model system and the real system. Different views are advanced regarding the nature of this relation. Peter Godfrey-Smith (2006) follows Giere (e.g. 1988) in talking of resemblance between model systems and the world, while Roman Frigg (2010b) speaks of a 'key' which specifies how facts about the model system are translated into claims about the real system.

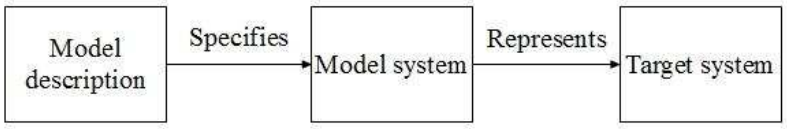


Figure 1: The indirect fictions view

The biggest challenge for the indirect fictions view concerns the ontology of model systems. After all, the nature of fictional characters is far from clear. Realists argue that, even if he is not a regular, flesh-and-blood detective, we must grant that Holmes exists in some sense. Holmes, along with Emma Bovary, Middle Earth and the rest, are therefore given a place in our ontology as fictional entities. Realists then offer different accounts of the nature of these entities. Meinong (1904/1960), for example, famously distinguishes being from existence. On this view, Holmes is an object possessing all the properties that we normally take him to have, such as being a detective and smoking a pipe; he simply lacks the property of existence. By contrast, antirealists, like Russell (1905/1956), aim to show how we can

---

interpretations of the ontology of modelling: the former takes representation to occur via a model system, and the latter does not.

understand fictional characters, and our talk about them, without granting the existence of fictional entities.

Proponents of the indirect fictions view have responded to this problem in a number of ways. Some look to existing theories of fictional characters. Thus, Roman Frigg (2010a, 2010b) aims to fill out the view by drawing on an existing antirealist theory of fiction. Ronald Giere (2009) suggests a different strategy. Although, in his earlier work, Giere takes models to be abstract objects, he has recently suggested that he too is willing to think of models as akin to fictional characters. But Giere argues that philosophers of science need not be too concerned with the question of exactly what such entities are. Peter Godfrey-Smith (2006) appears to endorse a similar strategy.

Let us now consider both of these responses in turn.

2.2. Antirealism and the Indirect Fictions View

Roman Frigg (2010a, 2010b) has proposed a version of the indirect fictions view that draws on an influential theory of fiction due to Kendall Walton (1990). On Frigg's view, model descriptions give rise to model systems, and these model systems are 'akin to characters and places in literary fiction' (2010b, 100). Frigg acknowledges, however, that without a theory of fictional characters 'explaining model systems in terms of fictional characters amounts to explaining the unclear with the obscure' (2010a, 256). It is for this reason that he looks to Walton's theory.

According to Walton, the text of a novel functions as a 'prop' in games of make-believe: when we read the text, we are supposed to imagine things according to certain rules (1990, chap. 2). Frigg offers an application of Walton's theory to scientists' model descriptions. On

Frigg's view, when we read the model description given by the Newtonian model of the solar system, for example,

> [w]e imagine the entity described in the description.... We understand the terms occurring in the description and we imagine an entity which has all the properties that the description specifies. The result of this process is the model system, the fictional scenario which is the vehicle of our reasoning: an imagined entity consisting of two spheres, etc. (2010b, 133; author's emphasis)

Frigg calls the relationship between the model description and the model system 'p-representation' (2010a, 264). When scientists want to represent a real system, like the sun and earth, they must establish a second representation relation between their model system and the world, which he calls 't-representation' (ibid.).

Frigg's aim, then, is to flesh out the indirect fiction view by drawing on an existing theory of fictional characters. The choice of Walton's theory for this task is a little surprising, however. The reason it is surprising is that Walton is an antirealist about fictional characters (1990, chap. 10). In Walton's view, works of fiction may seem to ask us to imagine things about people like Sherlock Holmes, and we may seem to be able to talk about them. But there simply are no such things, not even as Meinongian nonexistent entities. So if we were to understand model systems in the same way that Walton understands fictional characters then it seems that we would conclude that there are no model systems.

Frigg intends to follow Walton in his antirealism (2010a, 264; 2010b, 120). An antirealist stance on model systems is difficult to reconcile with Frigg's overall, indirect account of modelling, however. Model systems have a central place in that account: scientists use model systems to represent real systems (t-representation). According to Frigg's account of t-

representation, a model system X represents some real target system Y if and only if X denotes Y and 'X comes with a key K specifying how facts about X are to be translated into claims about *Y'* (2010b, 126). This might involve, for example, specifying 'object-to-object correlations', such as that 'the sphere with mass m_e in the model system corresponds to the earth and the sphere with mass m_s to the sun' (ibid., 134). Once we have specified such correlations

> we can then start translating facts about the model system into claims about the world. For instance, calculations reveal that the model-earth moves on an ellipse, and given that the model system is an ideal limit of the target we can infer that the real earth moves on a trajectory that is almost an ellipse. (ibid., 135)

If taken literally, however, all of these claims about t-representation would seem to be inconsistent with antirealism. If there are no model systems then there can be no facts about them and we cannot establish an object-to-object relation between model systems and the world. If there is no model-earth then it cannot move on an ellipse.

One way to reconcile Frigg's account with antirealism would be to offer some antirealist reinterpretation of what Frigg says about t-representation, which explains away the apparent commitment to fictional entities. If we were to take this route, however, all talk of using model systems to denote real systems, or of specifying object-to-object correlations between the two, would now be construed merely as a way of talking, rather than as offering an account of how modelling actually works.

Another option would be to abandon antirealism. Frigg suggests that he is open to this possibility (2010b, 113). And, in fact, Frigg's analysis of model systems differs from Walton's analysis of fiction at a number of points, and sometimes seems at odds with

antirealism. For example, he writes that 'the attribution of certain concrete properties to models … is explained as it being fictional that the model system possesses these properties' (2010b, 116; see also 2010a, 261). On Walton's theory, however, to say that it is fictional that the model system possesses certain properties is to say that we are to imagine that the model system possesses those properties. This would seem to conflict with antirealism: we cannot imagine things about model systems if there are none. However, if Frigg were to reject antirealism, and grant that we must posit fictional entities to serve as model systems, it seems that he would need to provide an account of what fictional entities are. And drawing on Walton's theory will not help to provide such an account.

2.3. Deferring the Problem

So the key challenge remains: can proponents of the indirect fictions view flesh out the comparison between model systems and fictional characters by providing a coherent account of what fictional characters are? As we saw earlier, however, some have argued that this challenge need not be met. In fact, they claim, worries about the ontology of fictional characters need not concern philosophers of science. For example, in his recent work, Ronald Giere grants that scientific models and fictional characters are ontologically 'on a par' (2009, 249). But he questions 'whether we, as philosophers of science interested in understanding the workings of modern science, need a deeper understanding of imaginative processes and of the objects produced by these processes' (ibid., 250). Peter Godfrey-Smith (2006) appears to endorse a similar attitude. Rather than defending any particular account of the ontology of fictional characters, he suggests that we might accept such objects as part of the 'folk ontology' of scientific modelling, even if in the long run we require an account of these objects 'for general philosophical reasons' (2006, 735).

I am sympathetic to this attitude. Later (Section 3.2) I will suggest that philosophers of science may indeed defer questions concerning fictional characters to philosophers of fiction. The important point to notice, however, is that this route is not open to those who defend the indirect fictions view. This view gives fictional characters a central place in modelling: on the indirect fiction view, scientists represent the world via fictional characters. To understand scientific representation we must therefore understand the relationship between a fictional character and the world. It is difficult to see how we could understand how such things represent without first understanding what they are.

For example, both Giere and Godfrey-Smith describe the relationship between model systems and the world in terms of similarities or resemblances between the two. If their accounts of the model-world relationship are to be taken literally then this will clearly place constraints on the account of fictional characters we can adopt: it must be a realist account, on which there are fictional entities and these entities can be said to possess properties such as mass or velocity. If we wanted to take a different view of fictional characters then all talk of similarity or resemblance between model systems and the world would have to be radically reinterpreted. If defenders of the indirect fictions view wish their accounts of scientific representation to aspire to truth, rather than being merely convenient stories, then it seems that they cannot leave fictional characters to philosophers of fiction.

3. A Direct Fictions View

3.1. Models and Fiction Revisited

As we have seen, some models (like the model of the sun and earth) represent real systems while others (like our predator-prey model) do not. The indirect fictions view suggests that we understand both in the same way: in each case, it is argued, the function of

the scientists' model description is to create a model system, which is akin to a fictional character. The only difference between the two sorts of cases concerns what the scientists do with the model system afterwards. When they model an actual system, scientists establish another representation relation between the model system and the world.

I think that these are the wrong parallels to draw between models and fiction. Rather than comparing all model descriptions to passages about fictional characters, I believe, we should distinguish carefully between cases where scientists model a real system and those where they do not. In the latter cases, model descriptions are like passages about fictional characters. In the former cases, however, scientists' model descriptions are more like works of historical fiction, that represent real people, places and events (for a similar suggestion, see Cartwright 1983, chap. 7). Consider the following passage, from Robert Graves' I, Claudius:

> Augustus assumed Antony's Eastern conquests as his own and became, as Livia had intended, the sole ruler of the Roman world. (Graves 1934/2006, 23)

As commonly understood, this passage is not about any fictional character, but about the real Emperor Augustus, as well as his wife Livia, Mark Antony and the Roman Empire.[2] According to Walton, for example, when we read fiction that uses the names of well-known figures like Augustus, the names take their usual referents (1990, chap. 3). On this view, a novel like I, Claudius represents real people, places and events, by asking us to imagine propositions about them. Some of these propositions are true, such as that Augustus defeated

[2] Not all will agree with this interpretation, of course. Fortunately, we need not enter into that debate here. (For a helpful review, see Friend 2007.)

Mark Antony. Others are, it seems, entirely fabricated by Graves and so probably false, such as that Augustus was manipulated by the scheming Livia.

This analysis of historical fiction suggests a better way to understand models of real systems. Recall Frigg's discussion of the solar system model. On his view, when we read the scientists' model description we first imagine an entity, the model system, which has all the properties given in the description. It is only in the 'next step', that we 'connect our model to the target-system' (2010b, 134), by specifying that the smaller sphere in the model system corresponds to the earth, the larger sphere to the sun, and so on. And yet it is surely more natural to regard the model description as asking us to imagine things about the sun and earth themselves. Frigg himself writes that the description 'tells us to regard the earth and sun as ideal homogeneous spheres' (ibid., 133), for example. Why not avoid excessive reconstruction and take the description at its word, as asking us to imagine things about the (actual) earth and the (actual) sun? Specifically, we are asked to imagine that the sun and earth are perfect spheres with certain masses, that they interact only with each other, and so on. Some of this is true (e.g. that the earth and sun are massive bodies) while some is known to be false (e.g. that they interact only with each other).

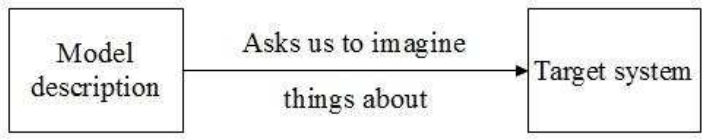


Figure 2: A direct fictions view

In place of the indirect fictions view, then, I propose a direct account (figure 2). When scientists model a real system they ask us to imagine things about that system directly, not via any fictional model system. As we saw in Section 1, sometimes the same model description

may be used in different ways. We might first explore the properties of the simple harmonic oscillator without having any real system in mind, and only later use it to understand the motion of the pendulum in the grandfather clock. According to the indirect view, when we apply the model to the pendulum, we do so by comparing our imaginary model system with the real pendulum. My suggestion is that another, more plausible, interpretation remains open to us. When we apply our model to the pendulum we simply imagine that the pendulum satisfies our model description. That is, we imagine that the pendulum is a point mass, that the force exerted on it is proportional to its displacement, and so on.

The point being made here thus involves drawing a distinction between two different sorts of imaginings. Sometimes, we imagine people, places and objects that do not exist, like Sherlock Holmes or imaginary populations of predators and prey. Sometimes, however, we imagine things about real objects or people in the world, as when I imagine that the walls in my flat are painted a different colour, or that I play for Derby County. The mistake made by proponents of the indirect view, I believe, is to assume that all cases of modelling involve cases of the first sort of imagining. It is true that scientists sometimes conjure up imagined systems, just as novelists create fictional characters. But we need not assume that, when the scientist comes to represent the world, she must somehow use these imagined systems to do so. Another option remains open: the scientist may simply imagine things about the world.

3.2. Avoiding Fictional Characters

The direct view allows us to leave problems with fictional characters to philosophers of fiction. Recall that this deferral strategy is not open to the indirect view because, on that view, when scientists represent the world they do so via fictional characters. As a result, our account of scientific representation becomes dependent upon which view of the ontology of fictional characters we adopt. This is not the case on the direct account. On the view I have

proposed, when scientists represent the world they do so by imagining propositions about it, not via a fictional character. Problems with fictional characters do still arise, but only for models that do not represent any real system, like our predator-prey model. And philosophers of science may legitimately defer these problems to philosophers of fiction. All that matters in these cases is that scientists are able to imagine things about objects that do not exist. Nobody doubts that we have this ability; the debate concerns how we are to explain it. And nothing in my account hinges on the outcome of this debate.

When scientists do not model a real system, then, I suggest that we remain neutral: perhaps we will need fictional entities to make sense of model descriptions, or perhaps not. Where scientists model a real system, however, we can be clear: there is no need to posit entities that satisfy the scientists' model descriptions. The model description asks us to imagine propositions about a real system, and many of these propositions are false. But nothing in this requires us to posit any fictional entity.

As we have seen, however, scientists often talk as if there were an object that satisfies their model description. How can the direct account make sense of this? One answer is suggested by Adam Toon (2010, 2012). Toon also draws on Walton's theory of fiction, but the main ideas behind his analysis may be summarised briefly here. When scientists talk about theoretical models as objects, Toon suggests, we should not take this talk too seriously. Instead, they are pretending, 'going along with' the model in order to tell us what we are to imagine. For example, suppose we say that in the model the sun and earth are isolated from the other planets. When we say this we are not describing any abstract or fictional object; we are simply saying that the model tells us to imagine that the sun and earth are isolated from the other planets.

Toon's analysis also suggests a way in which the direct account might explain how it is that we can learn about a model. Our initial model description asks us to imagine that certain assumptions are true of the sun and earth, such as that they are perfect spheres and that the force between them obeys Newton's law of gravitation. If we accept these initial assumptions, however, we are also required to imagine further things, which follow from those assumptions. For example, we are to imagine that the earth moves in an ellipse, since this follows from the equation of motion that we write down. That the earth moves in an ellipse is therefore part of the content of the model, even though this was not specified explicitly in the model description. On this view, then, learning about a model is not a matter of discovering facts about an abstract or fictional object. Instead, we learn about a model by exploring the intricate web of imaginings which it prescribes.

4. Conclusion

Parallels with fiction offer useful tools for understanding scientific models. But we should be careful what parallels we draw. Comparing all model descriptions to passages about fictional characters yields an implausible interpretation of what scientists are doing when they model a real system, and leads us to longstanding disputes over the nature of fictional characters. A more plausible approach looks to fiction about real people, places and events. On this view, when scientists model a real system, they represent that system directly by asking us to imagine it in a certain way, and not via any fictional character. As a result, philosophers of science may leave problems with Sherlock Holmes to philosophers of fiction.

References

Cartwright, Nancy. 1983. How the Laws of Physics Lie. Oxford: Clarendon.

Conan Doyle, Arthur. 1902/2003. The Hound of the Baskervilles. London: Penguin.

Contessa, Gabriele. 2010. "Scientific Models as Fictional Objects." Synthese 172(2): 215-229.

Friend, Stacie. 2007. "Fictional Characters." Philosophy Compass 2: 141-156.

Frigg, Roman. 2010a. "Models and Fiction." Synthese 172(2): 251-268.

———. 2010b. "Fiction and Scientific Representation." In Beyond Mimesis and Convention: Representation in Art and Science, ed. Roman Frigg and Matthew Hunter, 97-138. Dordrecht: Springer.

Giere, Ronald. 1988. Explaining Science. Chicago: University of Chicago Press.

———. 2009. "Why Scientific Models Should Not Be Regarded as Works of Fiction." In Fictions in Science: Philosophical Essays on Modeling and Idealization, ed. Mauricio Suárez, 248-258. London: Routledge.

Godfrey-Smith, Peter. 2006. "The Strategy of Model-Based Science." Biology and Philosophy 21: 725-740.

Graves, Robert. 1934/2006. I, Claudius. London: Penguin.

Hughes, R.I.G. 1997. "Models and Representation." Philosophy of Science 64: S325-S336.

Meinong, Alexius. 1904/1960. “The Theory of Objects.” In Realism and the Background of Phenomenology, ed. Roderick M. Chisholm, 76-117. New York: Free Press.

Russell, Bertrand. 1905/1956. “On Denoting.” In Logic and Knowledge, ed. Robert C. Marsh, 41-56. London: George Allen and Unwin.

Suárez, Mauricio, ed. 2009. Fictions in Science: Philosophical Essays on Modeling and Idealization. London: Routledge.

Thomson-Jones, Martin. 2007. “Missing Systems and the Face Value Practice.” Unpublished manuscript, retrieved from http://philsci-archive.pitt.edu/archive/00003519.

———. 2010. “Missing Systems and the Face Value Practice.” Synthese 172(2): 283-299.

Toon, Adam. 2010. “The Ontology of Theoretical Modelling: Models as Make-Believe.” Synthese 172(2): 301–315.

Toon, Adam. 2012. Models as Make-Believe: Imagination, Fiction and Scientific Representation. Basingstoke: Palgrave Macmillan.

Walton, Kendall. 1990. Mimesis as Make-Believe. Cambridge, Massachusetts: Harvard University Press.

Weisberg, Michael. 2007. “Who is a Modeler?” British Journal for Philosophy of Science, 58(2): 207-233.

1 March 2012

Causal relations and explanatory strategies in physics

Andrew Wayne

Draft – please do not cite

Word count: 4834 words

Abstract

Many philosophers now regard causal approaches to explanation as highly promising, even in physics. This is due in large part to James Woodward's influential argument that a wide range of explanations (including explanations in physics) are causal, based on his interventionist approach to causation. This article focuses on explanations, widespread in physics, involving highly idealized models. These explanations are not causal, yet they do not fall under any of the types of non-causal explanation Woodward describes. I argue that causal explanation is simply not as widespread or important in physics as Woodward and others maintain.

1. Introduction

Many philosophers now regard causal approaches to explanation as highly promising, even in physics. In part this is because the major alternative, deductivist approaches to explanation, have fallen on hard times (Hempel 1965; Kitcher 1989). Problems of explanatory irrelevance and explanatory asymmetry (recall hexing spells and flagpoles) have motivated many to pay more attention to the role of causation in explanation. Preeminent among recent work on causal explanation is James Woodward's influential argument that a wide range of explanations, including explanations in physics, are causal explanations, based on his interventionist approach to causation (Woodward 2003; Woodward 2007). After reviewing Woodward's approach (Section 2), this paper argues that causal relations are insufficient for explanation because they do not account for the key feature of explanatory integration in physics (Section 3). Further, causal relations are unnecessary for explanations, widespread in physics, involving highly idealized models. These explanations are not causal, yet they do not fall under any of the types of non-causal explanation Woodward describes (Section 4). This constitutes a significant limitation on the scope of causal explanation in

physics that neither Woodward nor any other proponent of causal explanation has recognized. Causal explanation is simply not as widespread or important in physics as Woodward and others—such as Wesley Salmon, Phil Dowe and Michael Strevens—maintain (Salmon 1984; Dowe 2000; Strevens 2008).

2. Woodward on causal explanation

For Woodward, causal relations are captured in counterfactual claims about what would happen to an effect Y if an intervention on another variable (or set of variables) *X* were to occur. Causal explanations in turn appeal to these "interventionist" counterfactual dependencies. Woodward is clear that his account of causation is non-reductive, in the sense that it does not aim to give an account of causation exclusively in non-causal terms. Explanation is also non-reductive, for Woodward. He allows that not all causal explanations need be in terms of fundamental physics, and indeed that fundamental physics is an area in which explanations seem to be predominantly non-causal. He emphasizes that macro causal claims can often be more explanatory than causal claims about their micro realizers, and that these macro causal claims can be explanatory while offering only an approximate description of the relevant features of the target physical system.

Consider an explanandum consisting of the statement that some variable *Y* takes a particular value. For Woodward,

(1) [A] successful [causal] explanation will involve a generalization G [in the explanans] and explanans variable(s) X such that G correctly describes how the value of Y would change under interventions that produce a range of different values of X in different background circumstances (2003, 203).

What makes the causal generalization G explanatory is that it can answer a relevant range of "what-if-things-had-been-different" questions, and it does this by supporting the correct counterfactuals about what would happen under scientifically relevant interventions on the explanans variable X. To do this, G must be invariant (roughly, describe the same sort of dependence of Y on the X) under the relevant range of interventions and in a range of relevant background conditions. Unlike deductivist approaches, successful explanations are not just nomologically sufficient, that is, they cannot just subsume the explanandum under a regularity and thereby show it is to be expected given the truth of the statements in the explanans. Rather, they must also describe relevant dependency relations—they must show how this explanandum would change if the intervention or background conditions were to change. Explanation locates the explanandum within a space of relevant alternative possible explananda.

We have seen that on Woodward's account, causal explanation requires counterfactuals describing possible interventions and possible covariation in changes in the values of

variables, and a notion of scientifically relevant possibility guiding the selection of interventions, dependencies and alternative possible explananda. The other key component of his account, of course, is an account of causal relations, including the cause-effect relation between variable X in the explanans and Y in the explanandum. For Woodward, if some intervention on X produces a change in the value of Y, then X is a token direct cause of Y. Roughly speaking,

(2) An *intervention I* is a hypothetical experimental manipulation on X such that,
(i) *I* causes X,
(ii) *I* changes the value of X in such a way that the value of X does not depend on the values of any other variables that cause X, and
(iii) *I* changes the value of X in such a way that if any change occurs in the value of Y, it occurs only as the result of the change in X and not from some other source.

(See Woodward 2003, 98-107 for a more detailed account.) Woodward's notion of intervention is not limited to what humans can actually do with physical systems. Rather, it is defined in terms of possible or hypothetical manipulations of values of variables within a model.

Woodward rightly emphasizes that only some changes in the explanans and only some contrasts between the explanandum and its alternatives are of causal and hence explanatory relevance. As he puts it, "It is also true that if a large meteor had struck my office just as I was typing these words, I would not have typed them, but again, we are reluctant to accept the failure of the meteor to strike as part of the explanation for my writing what I did" (2003, 226). The problem here is not that causal omissions can never figure in genuine explanations—Woodward is clear that sometimes they can—but rather that in this context a meteor intervention is not what Woodward dubs a "serious possibility." Scientists approach empirical phenomena with a large stock of shared beliefs about which of the interventions or dependency relations are potentially causally and explanatorily relevant, and which alternatives to the explanandum are relevant as well. Woodward is clear that what counts as a causal factor is relative to a particular choice of variables and also to a particular range of values of these variables (Woodward 2003, 55-56). Different models—in Woodward's terms, different sets of structural equations, variables and directed graphs—result in a different set of causes and hence a different explanation.

So far explanation, causation and intervention have been defined in terms of statements about variables, values and dependency relations within a model. But not every transformation or modification one can perform on a model corresponds to a hypothetical manipulation on the physical system itself (in Woodward's sense), and only those that do so correspond can underwrite causal claims. Causation requires that the values and dependency relations of variables in the model represent physical features of the target system. As Woodward puts it, successful causal explanation requires that the statements (about

counterfactuals, dependency relations, values of variables, causal relations and so on) in the explanandum and in the explanans be true or approximately true of the target system (2003, 203). Without the truth or approximate truth of the explanandum, it fails to be an explanation of any physical phenomenon at all. Without the truth or approximate truth of the explanans, the statements about the model simply cannot describe any real causal relations in the target system.

For example, the period of a pendulum may be approximately derived and explained in terms of its length, in a fixed gravitational field, by appealing to counterfactual claims about the behaviour of an idealized pendulum model satisfying Galileo's pendulum law. The law states that the period of a pendulum is proportional the square root of its length:

(3) $T \alpha \sqrt{l}$

The relevant counterfactual claim is: if the length l were increased to l^*, in a fixed gravitational field, then the period T of the model pendulum would have increased to T^*, in accordance with (3). However, the model does *not* support an explanation of the length of the pendulum in terms of its period, because the relevant interventionist counterfactual is false of the model: it is false that if the period were increased to T^*—for instance by moving the pendulum to a weaker gravitational field—the length of the pendulum would have changed. Woodward uses this example to illustrate how his causal model of explanation solves the problem of explanatory asymmetry that bedevils deductivist approaches (2003, 197). For our purposes, the important point is that the interventionist counterfactual doing the explanatory work (and described in the explanans) is true of the model and is also approximately true of the target system. For Woodward, the fact that the dependency relations in the model approximate "what the real dependency relations in the world actually are" is fundamental to his account of causal explanation (Woodward 2003, 202).

3. Causal relations are insufficient for explanation

I contend that a consequence of Woodward's account is that causal relations are insufficient for explanation in physics, and in two steps. First, some causal derivations fail to be explanatory. They may satisfy (1) and (2) above, and they may have significant predictive or heuristic value, but they do not explain. Second, where a causal derivation is explanatory, it is never merely by virtue of satisfying (1) and (2); rather, explanation requires that the causal story be integrated with a global model of broad scope and explanatory power.

According to Woodward, what makes the causal generalization G in (1) explanatory is that it answers "what-if-things-had-been-different" questions, and it does this by supporting the correct counterfactuals about what would happen under interventions. Consider Woodward's example of the explanation of the period of a pendulum, but this time prior to Galileo's theoretical advances. Taking liberties with the actual historical order of events,

imagine (counterfactually) that Galileo had conducted his years of painstaking experimental observations of pendulums first, in advance of any other work on his new science of mechanics. Had he arrived at his pendulum law (3) and his idealized pendulum model this way, we would be inclined to say that his argument deriving the period of a pendulum is not explanatory. The pendulum model on its own supports a relevant and approximately correct set of counterfactual claims about interventions on a physical pendulum. Nonetheless, it would be merely a phenomenological or data model, as contemporary physicists would put it. It fits a given set of data well, and it may describe the correct dependency relations in an isolated model, but fails to connect with other, more global models. These sorts of models may have predictive and heuristic power, but they do not underwrite explanations in physics.

Unfortunately, Woodward's account yields the result that many phenomenological models do come out as explanatory, and this cannot be right. Woodward posits a base threshold of explanatoriness, above which stands a continuum running from less deep or good explanations to deeper and better ones (2003, 368). The worry is that (1) and (2) set the threshold very low indeed: generalizations that are invariant under any intervention at all exceed the threshold because they answer a "what-if-things-had-been-different" question (2003, 369). So Woodward would certainly view the counterfactual Galileo's standalone pendulum model as underwriting a bona fide explanation of the period of the pendulum. But we have good reason to maintain that it does not, nor do the plethora of other phenomenological models in physics that capture some of the dependency relations in their target physical systems.

As a matter of historical fact, the pendulum law is significant for Galileo precisely because it is a key step in his route to the fundamental laws of his new science of mechanics. Galileo measured the elapsed time of an object's vertical fall over a distance equal to the length of the pendulum, for various pendulum lengths (Drake 1989, xxvii). He obtained a constant ratio of free-fall times to time for the pendulum to swing to vertical. With the pendulum law and that ratio, Galileo could calculate the times for other distances of free-fall and then, removing pendulums entirely from the calculation, write down his famous law of motion: that all objects fall at the same rate, regardless of their composition or mass, and that objects starting at rest accelerate uniformly as they fall, i.e. their speed is proportional to the square of the elapsed time of fall. He found the law fit well his previous measurements of descents along inclined planes.

This suggests that the idealized model pendulum gets its explanatory power by its integration into Galileo's new science of mechanics. In this case, it is integration of a particularly simple sort: Galileo took his pendulum law to follow from his more general law of free fall, and the idealized model pendulum is simply a special case of a more general model covering falling objects in general. Newton's subsequent achievement was greatly to increase this integration by explaining the motions of bodies in terms of the forces acting on

them and providing a unified framework for all gravitational systems. The important point for our purposes is that it is not sufficient that the idealized pendulum model approximate the correct dependency relations in a physical pendulum for it to be explanatory.

Woodward does say that successful causal explanation must include *relevant* dependency relations and answer a *relevant* range of "what-if-things-had-been-different" questions, and that scientists share an understanding of which interventions and which dependency relations are *explanatorily relevant*. Woodward seems to recognize that merely describing local causal relations is not sufficient for explanation, while perhaps not fully appreciating the consequences for the role of causation in explanation. The challenge is not to rule out an explanatory role for the absence of falling meteors. Rather, the challenge is to underwrite the explanatory role of dependency relations in the local pendulum model. And this can be done only in the context of a wider integration with a global model in physics—here Galilean (or even better Newtonian) mechanics.

The point is not just that some causal derivations satisfying (1) and (2) fail to be explanatory, as in the contrary-to-historical-fact Galilean account of the pendulum. It is also that no causal derivation is explanatory merely by virtue of satisfying (1) and (2). This is because what makes the dependency relations described in the explanans relevant (*i.e.*, explanatorily relevant) is the integration of the local model described in the explanans with a global model of broad scope and explanatory power. Without such integration, the local model will generally fail to be explanatory, no matter how accurately it represents causal relations in the target physical system. And as we shall now see, with such integration the local model will generally be explanatory—even if it fails to represents any causal relations in the target physical system.

4. Causal relations are unnecessary for explanation

Woodward allows that not all explanations in physics need be causal and notes that fundamental physics is an area in which explanations seem to be predominantly non-causal. What Woodward has in mind, in these and other sorts of physics explanations he calls non-causal, are cases in which the notion of an intervention on a physical system is incoherent or inapplicable. This includes global applications of fundamental physics to the whole universe or to large portions of it, where the notion of a local intervention is inapplicable (2007, 91); explanations that appeal to alternative situations not plausibly characterized as an intervention, e.g., altering the dimensionality of space-time (2003, 220); and situations that lack the invariance or stability properties needed to define an intervention on the system (2007, 77). These sorts of cases, however, are merely the tip of a very large iceberg of non-causal explanation in physics.

The issue is that, aside from explanations in textbooks (from which Woodward's examples seem to be drawn), much of the explanatory practice in physics does not fit

Woodward's characterization. These are cases in which the idealized models that underwrite putative explanations are largely non-representative of target physical systems. So while they approximately model the explanandum behaviour, they do not approximate aspects of the physical system described in the explanans. Moreover, these models are not corrigible, in the sense that they cannot be refined in a theoretically justified, non-ad hoc way to bring them in closer agreement with the target system. The point is that these are cases of explanation in which physicists view the scientifically relevant claims about interventions and systematic patterns of dependency relations that figure in a potential explanans to be statements about a highly idealized model, statements that are not even approximately true of the target system containing the phenomenon to be explained. If the explanatory practice of contemporary physics is taken seriously, there are highly idealized models of significant explanatory value.

Valuable work has been done by philosophers of physics on the possible explanatory roles of highly idealized models (Rueger 2001; Batterman 2002; Bokulich 2008; Batterman 2010; Bokulich 2011). Alisa Bokulich, for instance, has argued that "fictional models" can be explanatory if they meet certain conditions. Bokulich focuses on semi-classical models, which mix classical and quantum features. These models are known not to represent successfully the physical system because, for example, they include quantum particles following definite classical trajectories. The earliest and most well-known of these models is Niels Bohr's model of the hydrogen atom. As Bokulich puts it, "I want to defend the view that despite being a fiction, Bohr's model of the atom does in fact explain the spectrum of hydrogen" (Bokulich 2011, 42). Robert Batterman is interested in how highly idealized models explain the universality of structural features, such as the common characteristic shape of droplets at breakup when water drops fall from a dripping faucet.

> We can explain and understand (for large scales) why a given drop shape at breakup occurs and why it is to be expected. The answer depends essentially upon an appeal to the existence of a genuine singularity developing in the equations of motion in a finite time. It is because of this singularity that there is a decoupling of the breakup behaviour (characterized by the scaling solution) from the larger length scales such as those of the faucet diameter. Without a singularity, there is no scaling or similarity solution. Thus, the virtue of the hydrodynamic singularity is that it allows for the explanation of such universal behaviour. The very break-down of the continuum equations enables us to provide an explanation of universality (Batterman 2009, 442-443).

Asymptotic analyses that systematically abstract away from micro details enable idealized models to explain underlying structural or universal features. Batterman calls these "asymptotic explanations" (Batterman 2002, Ch. 4).

One option for Woodward and other proponents of causal explanation is simply to reject any role for highly idealized models in explanation. These are putative explanations that fail to meet Woodward's requirement for causal explanation, nor do they fall under his class of

non-causal explanations in physics. These models are simply highly inaccurate representations of the physical world. One could argue that highly idealized asymptotic and semi-classical models have great heuristic and predictive value, but do not underwrite explanations. They can play no part in underwriting the true causal premises needed in an acceptable explanation. In my view, this kind of wholesale rejection of any role for highly idealized models in explanation would be a mistake. A closer look reveals a more nuanced and complex set of considerations.

In the case of the Bohr model and other semi-classical models, there is no consensus among physicists that these models are explanatory, and rightly so. Clearly, their explanatory merits need to be examined on a case-by-case basis. At the very least, we have good reason to be skeptical that the Bohr model of the atom has any explanatory value, especially in light of the quite impressive explanations of the hydrogen spectrum given in terms of relativistic quantum theory.

The situation with respect to asymptotic models is somewhat different. On the one hand, a case can be made that at least one of these models may be eliminated (in principle at least) in scientific explanation (Redhead 2004; Belot 2005). On the other hand, these sorts of models are used widely and are regarded as underwriting among the best explanations on offer in physics today. In addition to analyzing the use of asymptotic models to explain drop formation in hydrodynamics, Batterman has explored the use of asymptotic models to explain critical phenomena in thermodynamics and to explain the rainbow in catastrophe optics (Batterman 2002). Similar sorts of highly idealized, asymptotic models are accepted as explanatory in many areas of physics beyond those that are the focus of Batterman (and his critics). For instance, these sorts of models are taken to underwrite explanations of a wide variety of non-linear dynamical systems, from a damped, driven oscillator model of the human heart to gravitational waves ([self-reference omitted]).

The gravitational waves case is particularly interesting. Physicists take themselves to have explained gravitational waves using Einstein's General Theory of Relativity (GTR). However, even in the simplest models of binary systems that produce gravitational waves, the Einstein Field Equations (the equations of GTR) cannot be solved directly. The reason is that these are a set of coupled, nonlinear equations governing the relation between the distribution of matter and energy in the universe and the curvature of space-time (of which gravitational waves are one feature). An attempt to solve the Einstein Field Equations directly by applying regular perturbation methods results in divergences (infinities) in values for the properties of gravitational waves observable from earth. So physics takes what is by now a familiar strategy: replace the intractable original problem with a tractable one, called the post-Newtonian approximation, that makes essential use of singular perturbation theory and asymptotic models. The empirical results are predictions and explanations of gravitational wave phenomena. These phenomena have not been observed (at the time of

writing), but a handful of large gravitational-wave detectors should soon reach sensitivities high enough for direct detection of gravitational waves (Pitkin, Reid et al. 2011; [self-reference omitted]).

We have good reason to accept, at least provisionally, explanations in physics based on highly idealized models. However, I am not claiming to have presented a conclusive argument for doing so. Obviously, much work remains to be done. Further analysis of the details of Bokulich's and Batterman's examples is needed, and vastly more cases of putative explanation via highly idealized models in physics need to be examined in detail. The question that needs to be asked of each case is: does explanation of a phenomenon ineliminably require appeal to a highly idealized model in this case? Nor am I claiming that "model explanation" or "asymptotic explanation" are adequate normative accounts of explanation in physics that can underwrite this sort of explanatory practice. Rather, I am claiming that philosophers have good reasons to take seriously the fact that the explanatory practice of physics includes a large class of explanations based on highly idealized models, explanations that are clearly not causal on Woodward's (nor any other plausible) account. I should also note that rejecting these sorts of cases wholesale as explanatory failures has as a consequence that physicists are massively mistaken about the explanatory merits of their theories and about the scope of their understanding of the natural world. This runs counter to Woodward's own project of offering an account of explanation that has normative and descriptive elements in reflective equilibrium, an account "significantly constrained by prior usage, practice and paradigmatic examples" (2003, 8).

The best option is to accept these sorts of cases as explanatory and recognize that the explanations fall outside the scope of *causal* explanation in physics. We have seen how Woodward allows that explanations in physics may be noncausal where the notion of an intervention is incoherent or inapplicable. Explanations appealing to highly idealized models constitute a new way in which the notion of an intervention is inapplicable. In these explanations, the correct counterfactual dependencies between *I*, X and Y may well obtain such that Woodward's conditions (2)(ii) and (2)(iii) are satisfied. In other words, these cases fit very well Woodward's central idea that explanations include statements of counterfactual dependencies describing the results of a hypothetical manipulation of variables in a model. However, the explanation is not causal because (i) is surely false: *I* does not cause X, because the dependency relations in the model do not correspond to or represent—even in an approximative way—physical dependency relations in the target system. Choosing this option is to acknowledge that there is a distinct, large and important class of non-causal explanations that have not been recognized by Woodward, nor, I suggest, by other proponents of causal explanation in physics.

5. Conclusion

Recall that for Woodward, the notion of an intervention plays the crucial roles of underpinning both the *truth* and *explanatory relevance* of generalization G in the explanans of a successful causal explanation (1). In the context of physics, I have argued, "intervention" is simply not the right concept to play these roles. Even in cases where the notion of an intervention is coherent and applicable, it is not sufficient to meet the threshold of genuine explanatoriness in physics. As we have seen, what makes the dependency relations described in the explanans explanatorily relevant is the integration of the local model described in the explanans with a global model of broad scope and explanatory power. In other cases the notion of intervention is wholly unnecessary to underpin the truth of G, because G can be made true by facts about dependency relations in a model. These dependency relations are clearly not causal, because they are features of an idealized model that do not accurately represent corresponding features of the physical world.

Among the many virtues of Woodward's account of explanation are that it is explicitly model-based and that it makes explanation trace systematic patterns of dependencies rather than simply describing nomologically sufficient conditions. However, the argument given above that much successful explanation in physics involves highly idealized models counters Woodward's claim that many (non-fundamental) explanations in physics are causal. I suggest that the argument against Woodward's causal account tells equally strongly against other prominent defences of causal explanation in physics (e.g., Salmon 1984; Dowe 2000; Strevens 2008). There is good reason to believe that outside of textbook presentations, causal explanation is not as widespread in physics as its proponents have claimed. This point likely generalizes to other areas of science in which complex non-linear dynamical systems are modeled, such as biology and chemistry. These areas seem to have the same sorts of non-reductive explanations appealing to highly idealized, partially non-representative models. If this is right, causal concepts are not as useful in scientific explanation as many philosophers currently believe, and certainly causal theories of explanation are not as successful as the current consensus holds. Perhaps deductivist approaches to explanation merit renewed interest.

References

Batterman, Robert W. 2002. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford: Oxford University Press.

---. 2009. "Idealization and Modeling." *Synthese* **169**: 427-446.

---. 2010. "On the Explanatory Role of Mathematics in Empirical Science." *British Journal for the Philosophy of Science* **6**(1): 1-25.

Bokulich, Alisa. 2008. *Reexamining the Relationship between Classical and Quantum Mechanics: Beyond Reductionism and Pluralism*. Cambridge: Cambridge University Press.

---. 2011. "How Scientific Models Can Explain." *Synthese* **180**: 33-45.

Dowe, Philip. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Drake, Stillman. 1989. "Introduction." *Two New Sciences, Including Centers of Gravity and Force of Percussion*. Stillman Drake. Toronto: Wall & Thompson**:** i-xxxv.

Hempel, Carl Gustav. 1965. *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*. New York: Free Press.

Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." *Minnesota Studies in the Philosophy of Science, volume XIII*. Philip Kitcher and Wesley C. Salmon. Minneapolis: University of Minnesota Press**:** 410-506.

Pitkin, Matthew, Stuart Reid, et al. (2011) "Gravitational Wave Detection by Interferometry (Ground and Space)." Living Revies of Relativity **14**.

Redhead, Michael. 2004. "Asymptotic Reasoning." *Studies in History and Philosophy of Modern Physics, vol. 35, pt. B, no. 3, pp*: 527-530.

Rueger, Alexander. 2001. "Explanations at Multiple Levels." *Minds and Machines* **11**(4): 503-520.

Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, N.J.: Princeton University Press.

Strevens, Michael. 2008. *Depth : An Account of Scientific Explanation*. Cambridge, Mass.: Harvard University Press.

Woodward, James. 2003. *Making Things Happen : A Theory of Causal Explanation*. Oxford: Oxford University Press.

---. 2007. "Causation with a Human Face." *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Huw Price and Richard Corry. Oxford: Clarendon Press**:** 66-105.

An Epistemology of Causal Inference from Experiment

Karen R. Zwier

Abstract

The manipulationist account of causation provides a conceptual analysis of cause-effect relationships in terms of hypothetical experiments. It also explains why and how experiments are used for the empirical testing of causal claims. This paper attempts to apply the manipulationist account of causation to a broader range of experiments—a range that extends beyond experiments explicitly designed for the testing of causal claims. I aim to show (1) that the set of causal inferences afforded by an experiment is determined solely on the basis of contrasting case structures that I call “experimental series”, and (2) that the conditions that suffice for causal inference obtain quite commonly, even among “ordinary” experiments that are not explicitly designed for the testing of causal claims.

1. Introduction. The manipulationist account of causation, exemplified especially in the work of Woodward (2003), is a powerful and interesting explication of the meaning of causal claims. The account is intended as a conceptual clarification of what it is to be a causal relationship, and it provides this clarification by making reference to hypothetical experiments and ideal interventions. And since, according to the account, hypothetical experiments are embedded in the very content of causal claims, it requires only a small logical step to explain the role of experimentation in the empirical investigation of causal claims.

No one can deny that *some* scientists intend to test causal claims, and that they design and carry out experiments for the purpose. Does the type of fertilizer applied to potatoes affect crop yield? A scientist might perform an experiment by applying different types and quantities of fertilizer and comparing the resulting yield. Does a certain drug improve prognosis for patients with a certain condition? A group of scientists might perform a series of randomized, double-blind trials to find out. The manipulationist account certainly seems to be applicable for analyzing the success or failure of causal inference in experiments such as these. However, it is not quite as easy to see if—or how—the manipulationist account might apply to experiments that are *not* explicitly designed or carried out for the purpose of testing a causal claim.

Experiments in the physical sciences, in particular, rarely seem to be framed in terms of causal questions, at least not explicit ones. Consider an experiment aimed at measuring the boiling temperature of nitric acid at atmospheric pressure. Is such an experiment intended to test a causal claim? It certainly doesn't seem so, at least not at first glance. But could the experiment still afford causal inference, if we knew where to look and what assumptions to apply? I take the answer to this latter question to be non-obvious, and the goal of this paper is to make some progress toward an answer.

This paper attempts to apply the manipulationist account of causation to a broader range of experiments—a range that extends beyond the set of experiments that are explicitly designed for the testing of causal claims. I wish to include anything that we might naturally call an "experiment"—i.e., a scientific study in which the investigator deliberately sets up and/or intervenes on a system for the purpose of studying it.[1] I aim to show (1) that the set of causal inferences afforded by an experiment is determined solely on the basis of contrasting case structures that I call "experimental series", and (2) that the conditions that suffice for causal inference obtain quite commonly, even among "ordinary" experiments that are not explicitly designed for the testing of causal claims.

The implications of this point are potentially far-reaching. Even experiments not branded as "causal", including those carried out in the course of research in the physical sciences, can, under certain circumstances, afford causal inference. As a result, an experiment that meets certain criteria has the ability to furnish causal content even in those areas of science (e.g., fundamental physics) where causal content is less obvious.

2. The Manipulationist Account of Causation. I begin with a brief overview of the manipulationist account of causation. The manipulationist account, in its most basic form, is intended as an account of the *meaning* of causal claims. A meaningful causal claim must have an interpretation that refers to the result of some relevant hypothetical experiment. But what is the relevant hypothetical experiment for a given causal claim? Roughly, the idea is the following: for a causal claim such as "X causes Y", the hypothetical experiment under

[1] Purely observational studies (e.g., observing astronomical events through a telescope, analyzing retrospective health information, etc.) that involve no intervention or set-up on the part of the investigator will not be considered experiments for my purposes here.

consideration is one in which the variable or factor X is manipulated or changed in some way, and any corresponding change (or non-change) in Y is observed. According to the manipulationist account of causation, consideration of such an experiment is logically embedded in the very content of a well-formed causal claim, such that evaluation of the truth or falsity of the claim will be tied to an evaluation of whether or not a change in X would be seen if the experiment were to be performed.

We can state the idea more formally as a criterion for X to be considered a cause of Y:

> MANIPULATIONIST CAUSE: X is a *cause*[2] of Y iff, under some set of background conditions $\mathbf{BC} = \{BC_1, BC_2, \ldots, BC_n\}$ having values $\{bc_1, bc_2, \ldots, bc_n\}$, given some (possibly empty) set $\mathbf{S} = \{S_1, S_2, \ldots, S_m\}$ of variables other than X and Y that are held fixed at predetermined values $\{s_1, s_2, \ldots, s_m\}$, there is some ideal intervention[3] I on X that would change the value of Y.

[2]"Cause", as I use it here and throughout this paper, corresponds to Woodward's "type-level contributing cause". The criterion that I give here is a modified and simplified version of Woodward's **M**, which requires detailed knowledge of the path from X to Y (see Woodward (2003, 59)). In the context of my discussion here, I do not wish to assume that an evaluator of causal claims always has that knowledge, and so I give a criterion that does not require it. In addition, my criterion is intended to be more faithful to the implicit criterion for causation in the mind of an actual experimenter who is—implicitly or explicitly—testing a causal claim.

[3]The manipulationist account requires that an intervention variable have particular characteristics in relation to both X and Y and the larger system of variables being considered. For the purpose of brevity, I will not discuss these requirements here; see Woodward (2003).

According to the above criterion, a hypothetical experiment relevant to the evaluation of the claim "X is a cause of Y" is one in which we hold some (possibly empty) set of variables $\mathbf{S}$ fixed while intervening on X, and we observe any associated changes in the value of Y. The claim "X is a cause of Y" will be true if and only if changes would be observed in Y in the context of *some* hypothetical experiment defined by a specific $\mathbf{BC}$, $\mathbf{S}$, and I.

An important thing to note about this way of spelling out the meaning of a causal claim is that it makes use of a particular kind of counterfactual claim. In order to make sense of how an intervention on one variable, X, "makes a difference" to another variable, Y, we need to have some concept of what *would have happened* had the intervention on X not occurred. It is only by comparing the case in which the intervention is performed with our background understanding of what would have happened had the intervention not been performed (or had a different intervention been performed) that we get a sense of an effect.

A second thing to notice about the hypothetical experiment referenced by a causal claim is that it involves two different types of interactions with the experimental system. The values of the background condition variables in $\mathbf{BC}$ are observed, as is the value of Y. Nothing is done to directly force these variables to take on particular values. For X and for the set $\mathbf{S}$, however, interventions directly force these variables to take on certain values.[4] The distinction between *observing* the value of certain variables and *intervening* to set the value of others is absolutely central to the manipulationist account of causation. The character of the knowledge that we

[4]Experiments with a non-empty $\mathbf{S}$ will be multiple-intervention experiments intended for ruling out "unfaithfulness", as it is called in the causal modeling literature. In cases of unfaithfulness, observational data (and even some experimental data) can make it appear that two variables are independent of one another despite one being a cause of the other. See Spirtes et al. (2000, 13–14), Woodward (2003, 49–50), and Zhang and Spirtes (2008).

gain from observing a natural course of events in a system and that of the knowledge that we can gain from carefully designed interventions on that same system are essentially different. When we know from mere observation that certain values of X are associated with certain values of Y, this fact underdetermines the various types of causal connections that might exist between the two variables. Assuming that the correlation is not a spurious result of sample or selection bias, there are three different ways in which the variables might be causally connected: (i) X could be a cause of Y, (ii) Y could be a cause of X, and/or (iii) X and Y could share a common cause (or set of common causes). Interventions allow us to distinguish among these three types of causal connections (and their several combinations), because each kind of causal connection between X and Y would respond differently to interventions on X or Y.

3. From Hypothetical Experiment to Real Experiment. The conceptual tools and criteria discussed in the previous section serve the primary goal of the manipulationist account of causation: that of explicating and interpreting causal claims in terms of hypothetical experiments. Given a causal claim, these tools allow us to reconstruct the relevant hypothetical experiment embedded in the claim (or a set of relevant hypothetical experiments that reflect alternate interpretations of the claim).

Although the conceptual interpretation of causal claims is the primary goal of the manipulationist account, the manipulationist account of causation carries with it an important corollary for scientific practice. For those who wish not only to evaluate the content of a causal claim but moreover to test its truth, the manipulationist account can provide norms and recommendations for experimental testing. The truth or falsity of a causal claim can be empirically tested as long the hypothetical experiment embedded in the content of the claim

can be actually realized. Actual experiments intended to test a causal claim can—and should—be modeled on the hypothetical experiment suggested in the content of the causal claim.

Let us focus on how an actual experiment must be carried out if it is to test a causal claim:

> EXPERIMENTAL INSTANCE FOR TESTING THE CLAIM "X IS A CAUSE OF Y": Under some set of background conditions $\mathbf{BC} = \{BC_1, BC_2, \ldots, BC_n\}$ having values $\{bc_1, bc_2, \ldots, bc_n\}$, hold some set $\mathbf{S} = \{S_1, S_2, \ldots, S_m\}$ of variables other than X and Y fixed at values $\{s_1, s_2, \ldots, s_m\}$, perform an intervention I on X, and observe the value of Y.

The above operation, however, is only a single instance of an experiment and is insufficient for answering the question "Is X a cause of Y?" Recall that the hypothetical experiment embodied in the claim that X causes Y makes use of a contrast between two counterfactual states: the state of Y when X is manipulated in one way, and the state of Y if X *had been* manipulated in a different way (or not at all). But actual experiments provide us no access to such counterfactual knowledge.

The obvious way to estimate the results of counterfactual experimental instances is to test many instances of the experimental system under similar conditions and to use statistical analysis[5] to estimate the expected response of the system under different interventions. Let us define for this purpose an *experimental series*:

[5]Statistical analysis, as I intend it here, could be as simple as calculating a mean and standard deviation from the set of measured results, or could involve the application of much more sophisticated analysis techniques.

EXPERIMENTAL SERIES FOR TESTING THE CLAIM "X IS A CAUSE OF Y": A set of two or more experimental instances for testing the claim "X is a cause of Y" such that:

1. Every instance in the set has the same (or sufficiently similar) values for $\mathbf{BC}$ and $\mathbf{S}$; and
2. The set can be partitioned into two or more non-empty subsets such that every instance in each subset has the same value for the intervention I on X and no two instances falling into different subsets have the same value for the intervention I on X.

Observations made of the value of Y for each of the subsets described in item 2 above can be collated and used to generate a statistical estimate of the expected value of Y under the type of intervention used in that subset of experimental instances. If there is a significant difference in the expected values of Y for different subsets, then we may conclude that X is a cause of Y. If there is not a significant difference in the expected value of Y for different subsets, the conclusion must be more tentative. If a sufficient number of instances has been tested, we can legitimately conclude only that X is not a cause of Y under the particular circumstances of the experiment (where "circumstances" includes the background conditions $\mathbf{BC}$, the choice of $\mathbf{S}$ on which to perform secondary interventions, and the range of values of X that were effectively tested in the series). The possibility that X will manifest itself as a cause of Y under other circumstances remains open, but the likelihood of that possibility can be reduced by testing of other series with different values for $\mathbf{BC}$, different values for $\mathbf{S}$, and/or interventions testing differing ranges of values of X.

4. From Real Experiments to Causal Claims. We have already discussed the way in which a real experiment can approximate the hypothetical experiment embedded in a causal claim. Now I would like to turn our attention to experiments that are *not* explicitly concerned with causation or the testing of causal claims. When analyzing an experiment that was not designed for the purpose of testing causal claims, we simply seek to identify anything that could be properly described as an experimental series (on the definition given in the previous section).

Consider as an example an experiment performed by Gasparo Berti, which aimed to decide a philosophical controversy surrounding the possibility of a vacuum and test Galileo's predictions about the maximum height to which water could be raised by suction. The experiment was most likely carried out sometime in the years 1642–1643 in the company of several active participants in the scientific scene of Rome, including Raffaello Magiotti, Athanasius Kircher, and Niccolò Zucchi. A description of the experiment is found in a 1648 letter from Magiotti to Marin Mersenne. The following is an excerpt from the letter:

> In regard to the history of quicksilver, you may know that the many wells of Florence, which are cleaned each year by suction with siphons, gave Sig. Galileo the opportunity to observe the height of the attraction which was always the same, about 18 Tuscan *braccia*,[6] and that in every siphon or cylinder, no matter how wide or thin. This was the origin of his speculations on the subject in his work on the cohesion of solids.
>
> Later, Sig. Gasparo Berti, here in Rome, made a lead siphon that stretched about 22 *braccia* from his courtyard to his room, and was filled from above in the following way. First, leaving both valves open (D below and F above), vessel AG was filled with water. [See figure 1.] Then, after closing valve D, the water of vessel AGPM was poured out

[6]The *braccio* was equivalent to slightly more than half of a meter.

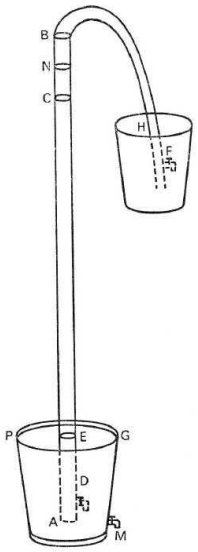


Figure 1: Diagram of Berti's experiment, included in Magiotti's letter to Mersenne

(through valve M), leaving the water inside the siphon at height AE. Later, making sure to keep vessel HF full,[7] the water AE was allowed to flow out through valve D, which (since valve F was already open and immersed in water) pulled the water from above and filled the whole siphon BA and the vessel AG. Finally, with vessel HF full and having closed valve F, and with vessel AG full (having first closed M) and D open, the water started to descend through the siphon, emptying the entire neck BF. The water continued to fall until reaching N and did not descend further, but almost always balanced itself [at N] when the experience was replicated. And it was possible to observe this very well, since part BC of the siphon was made of glass on purpose and the whole siphon was well glued and watertight. Sig. Berti believed that he could refute Sig. Galileo with this experience, saying that the length from N to A was more than 18

[7]This was presumably done by continuous refilling.

> *braccia*, but he should have seen that the piece of the siphon AE doesn't count, being immersed in the water of vessel AG; EN was 18 *braccia* exactly.
>
> I should not fail to mention one thing that gave me much to think about: while the water of the siphon was falling and the neck BF was emptying, an infinite number of tiny bubbles, like those in glasses and crystals, could be seen rising through the water inside the glass BC: this, without a doubt, was some stuff that went to refill where the air was missing. I could not convince myself that it was air because there was not enough air in the water in vessel AG to refill that space (besides, the space NBF could be made much larger and it would still refill). Nor could air have entered through pores or the welding of the siphon, for if it had, it would have eventually allowed the suspended water to fall. In fact, those bubbles have always remained in my mind: I can only explain my whole sentiment about them briefly like that.[8]

Besides Magiotti's letter, there are four other sources that describe Berti's experiment: two written by eyewitnesses Zucchi and Kircher, and two other secondary sources.[9] These other accounts all describe a similar and slightly more complex version of the experiment, which may have been a later modification. In this version, a glass globe was mounted on the siphon (see figure 2). The globe contained a bell attached to a magnetic device so that, once the purported vacuum was achieved inside the globe, the bell could be rung from outside by using another magnet.

The primary intention of the experiment, at least on Berti's part, appears to have been a desire to check (and perhaps refute) Galileo's prediction of 18 *braccia*. A secondary intention

[8]Translation mine. The manuscript of the letter is published in de Waard (1936, 178–181).

[9]de Waard (1936) contains relevant excerpts (in the original Latin) from all four sources.

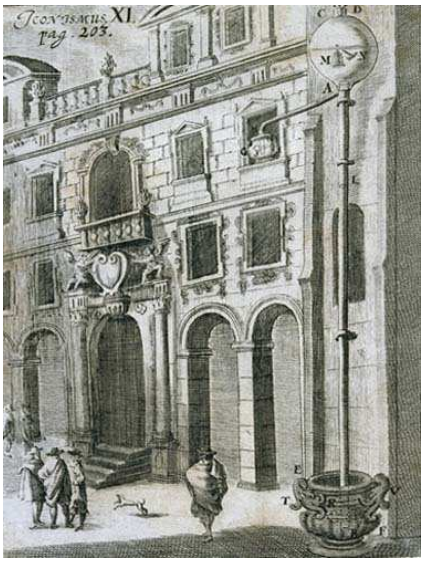


Figure 2: Engraving of a more complex version of Berti's experiment, reproduced in Schott (1664/1687, 203)

was to investigate the empty space itself: was it or was it not a vacuum? It is obvious from Magiotti's letter that this latter was a question of interest for him, and it was likely the most important question in the minds of the other participants as well; Zucchi and Kircher were both Jesuits who were convinced of the impossibility of the vacuum.

The addition of the bell in the more complex version of the experiment was suggested by Kircher and intended as an *experimentum crucis* to test the claim that the space in the globe was a vacuum. The space was found to transmit both light and magnetism, and the bell could indeed be heard when rung. These facts were enough to convince both Zucchi and Kircher, and perhaps also Berti, that the space was not a vacuum. Maignan, a friend of Berti's and a later commentator on the experiment, proposed the counter-opinion that the sound of the bell was being conducted by the bell's wooden support rather than by the space itself, and argued that the space was indeed a vacuum. It seems that Magiotti remained uncertain. Inasmuch as

the various participants walked away from the experiment with different views, the *experimentum crucis* was a failure.

Notice that the questions of interest for those performing and attending the Berti experiment were not causal questions; none of the writings explicitly mention a curiosity about the cause of the empty space, for example, nor is there any evidence of debate among the participants about what caused the elevation of the water to be 18 *braccia* rather than some other height. The questions posed and debated were, instead, factual questions and questions of interpretation about the phenomenon: How high did the water stand? Could there be any pores or imperfections in the device? Did the space transmit sound? Was the space a vacuum, or was it not?

Despite the lack of interest in causal questions on the part of those involved in the experiment, can causal conclusions can be drawn anyway? A first step toward deciding this question is to itemize the procedure described in the excerpt from Magiotti's letter and classify each step as an intervention component (I) or an observation component (O):

1. (I) Construct and set up the pipe and vessels in the configuration given in figure 1. Ensure that valve M is closed.
2. (I) Open valves D and F.
3. (I) Fill vessel AG with water.
4. (I) Open valve M.
5. (O) Observe that vessel AG empties. Water inside the siphon remains at height AE.
6. (I) Fill vessel HF with water.
7. (I) Open valve D and continue supplying HF with water.
8. (O) Observe that the water flows out through valve D and also flows from above to fill siphon.

9. (I) Close valve F and valve M.
10. (O) Observe that the water begins to descend down the siphon, emptying neck BF and falling until it reaches N.

Assuming a similar set-up for the more complex version of the experiment,[10] we might simply modify the first step and add several steps to the end of the procedure:

1*. (I) Construct pipe mounted with glass globe and internal magnet-bell apparatus. Arrange it and vessels in the configuration given in figure 2.

⋮

11. (O) Observe that light passes through the sphere.
12. (I) Move magnet around the exterior of the glass globe.
13. (O) Observe that the interior magnet moves in response to the exterior magnet's movement.
14. (O) Observe that sound can be heard from the bell inside the glass sphere.

It is interesting to notice that many—not just one—of the steps listed in the above procedures are interventions on the experimental system. Most of them serve only as steps toward the set-up of the apparatus. However, each can, in principle, be considered as an intervention in an experimental instance for testing a variety of causal claims; the variable X will be the thing intervened upon (for example, the intervention in step 4 is an intervention on whether or not valve M is open), the variable Y can be any observation that follows (for example, the observation in step 5 that vessel AG empties), and all other observations and

[10]Other accounts of the experiment describe a different procedure for filling the apparatus with water, but the difference in procedure is inconsequential for the analysis I offer below.

interventions involved in the experiment are considered either as observed background conditions in $\mathbf{BC}$ or auxiliary interventions in $\mathbf{S}$.

The question of whether or not the experiment affords causal inference amounts to the question of whether or not the various experimental instances that make up the experiment are part of an identifiable experimental *series*. Consider, for example, an experimental instance centered around the intervention in step 4 above. The variable X might represent the state of valve M (open or closed) and the variable Y might represent the state of the vessel AG (which can be empty or full, but is observed as empty in step 5). The set-up established in steps 1–3 and other background conditions surrounding the experiment could all be represented by the set $\mathbf{BC}$. Now, if we can identify at least one other experimental instance with the exact same values for $\mathbf{BC}$ but a different intervention on valve M, we will have identified an experimental series for testing the claim that the state of valve M is a cause of the vessel AG emptying. Berti's experiment does in fact provide such an experimental instance. Assuming that there is some time lapse between the execution of steps 3 and 4, we can consider as a second experimental instance the time period after steps 1–3 have been performed but before valve M has been opened. In this time period, vessel AG is observed to be full. Since there is a difference in the state of vessel AG between the experimental instance in which M is opened and the experimental instance in which M is not opened, we can conclude that the state of valve M is a cause of the state of vessel AG.

The observation-intervention pair considered in the example experimental series just given (i.e., a valve being opened and a vessel emptying) are such an ordinary matter of course that we do not tend to think of it as the basis for a causal conclusion that can be drawn from the experiment. That water only empties from a vessel that has some open outlet is a mundane fact that each person experiences so many times in life that it becomes an implicit piece of

causal knowledge. Still, inasmuch as the experiment establishes a contrast between performing and not performing an intervention (or alternatively, performing one type of intervention vs. performing a different type of intervention) and the corresponding difference in the observations made in each case, the experiment also affords the conclusion that one variable (the variable intervened upon) causes another (the variable observed to covary with the variable intervened upon).

But are there more substantial causal questions that could have been answered by the experiment in question? The interventions performed in the more complex version of the experiment, if compared to a relevant contrast case, could be interpreted as tests of causal questions. For example, when it is observed in step 11 that light passes through the spherical glass vessel, the implicit contrast case is whether or not light passes through the spherical glass vessel when it was originally filled with ordinary air. Presumably there were no noticeable differences between the appearance of images viewed through the vessel in the two cases. Likewise, we might compare the intervention in step 12 when it is performed in the context of the experimental set-up and when it is performed in a contrasting context (for example, with a column of water filling the siphon up to mark N, but not brought about through suction, so that the spherical glass vessel is filled with ordinary air).

The participants in the experiment were not, however, thinking in terms of these contrasting experimental instances. Even if they had been, they would have been unable to agree on a causal conclusion because they were unable to agree about what the interventions in the experiment had achieved. It is clear in Berti's experiment what the intervention is (or rather, what the sequence of interventions is: steps 1–4, 6–7, 9, 12) but what those interventions achieve was precisely the subject of debate. Some of the participants—the vacuists—thought that those interventions achieved a vacuum in the spherical vessel, while

others—the plenists—thought that the vessel was still filled with some sort of attenuated matter. If they had been able to agree, for example, that there was a vacuum in the vessel, then they might have been able to agree that ordinary air (as opposed to vacuum) was not a cause of the transmission of light or magnetism. In addition, they would have been able to reach a conclusion about the effect of the vacuum on the transmission of sound by noting any difference in the volume of the bell's ring in each case.

But there was no such agreement. Instead, some of the participants were already certain, prior to the experiment, that a vacuum could not transmit light or sound or magnetic phenomena. They took themselves to be certain of the causal relationships, and they attempted to test the presence or absence of the vacuum by the presence or absence of its purported effects. An experiment which could have been understood to test various causal claims instead used a prior confidence in those causal claims to test whether or not the cause factor was present. Even so, the actual theoretical use to which the experiment was originally put does not prevent anyone who is later informed of the details of the experiment from drawing causal conclusions.

5. Conclusion. Many experiments that are not designed for the purpose of causal inference will still afford causal inferences. The requirements I have placed on an experimental series for testing a causal claim will be found quite commonly in "ordinary" scientific experiments. We can see that this is true especially when we consider that, in cases where there is a time lapse between the set-up of the experiment and the intervention on the purported cause variable (if the time latency of the observed result is small in comparison to the time lapse), a comparison of observations made before and after the intervention is performed will usually correspond to an experimental series for testing if the variable intervened on is a cause of the

subsequent observation.

Interestingly, the fact that many "ordinary" experiments will afford causal inference means that any experimental science has a plentiful source of causal content. I see this unacknowledged point as significant to debates about whether or not there is causal content in fundamental physics.[11] In acknowledging the epistemic dependence of fundamental physics on experiment, we must also acknowledge at least a potential for causal content.

[11]For a set of papers in this debate, see the volume edited by Price and Corry (2007).

References

de Waard, Cornélius. 1936. *L'Expérience Barométrique: Ses Antécédents et Ses Explications*. Thouars: Imprimerie Nouvelle.

Price, Huw and Richard Corry, eds. 2007. *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Clarendon Press.

Schott, Gaspar. 1664/1687. *Technica Curiosa, Sive Mirabilia Artis*. Herbipol.: Jobus Hertz.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: The MIT Press.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Zhang, Jiji and Peter Spirtes. 2008. "Detection of Unfaithfulness and Robust Causal Inference." *Minds and Machines* 18 (2): 239–271.

Draft. Please do not quote without permission. Comments welcome.

Psychophysical Methods and the Evasion of Introspection

M. Chirimuuta
Dept. History & Philosophy of Science
1017 Cathedral of Learning
4200 Fifth Avenue
University of Pittsburgh, Pittsburgh, PA 15260

mac289@pitt.edu

Abstract

While introspective methods went out of favour with the decline of Titchener's analytic school, many important questions concern the rehabilitation of introspection in contemporary psychology. Hatfield (2005) rightly points out that introspective methods should not be confused with analytic ones, and goes on to describe their "ineliminable role" in perceptual psychology. Here I argue that certain methodological conventions within psychophysics reflect a continued uncertainty over appropriate use of subjects' perceptual observations and the reliability of their introspective judgements.

My first claim is that different psychophysical methods do not rely equally on the introspective capabilities of experimental subjects. I contrast "minimally-introspective" tasks with "introspection-heavy" ones. It is only in the latter, I argue, that introspection can be said to have a non-trivial role in the subjects' performance. My second claim is that my rough-and-ready distinction maps onto a number of important "dichotomies" in vision science (Kingdom and Prins 2009). Not coincidentally, the introspection-heavy categorisation captures many of the tasks typically considered less able to yield useful information regarding the processes underlying visual sensation.

1. Introduction

Recent work on introspection in psychology has been careful to separate the specific commitments of Titchener's analytical school from the discussion of introspection more generally. For example, Hatfield (2005) defines introspection broadly as, "a mental state or activity in or through which persons are aware of properties or aspects of their own conscious experience" (p.260). He later defines introspection as, "deliberate and immediate attention to certain aspects of phenomenal experience," arguing that, "it continues to be used as a source of evidence in perceptual and cognitive psychology" (p.279). In this paper I will challenge the appropriateness of Hatfield's definitions in the branch of perceptual psychology known as psychophysics[1], offering an alternative account.

[1] Psychophysics is defined by Gescheider (1997) as "the scientific study of the relation between stimulus and sensation." The disciplinary demarcation between psychophysics and perceptual

Draft. Please do not quote without permission. Comments welcome.

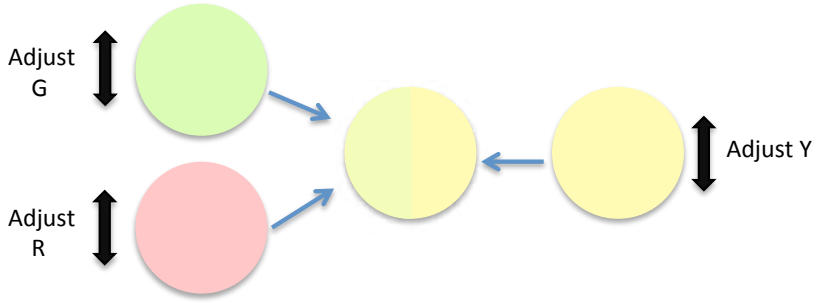


Figure 1
A Metameric Match Experiment. The subject is asked to adjust the intensities of R, G and Y monochromatic lights so that the yellows are indistinguishable.

Hatfield discusses the psychophysical task of metameric colour matching (see fig. 1) as one example of a perceptual experiment reliant on introspection (p.278). However, it is reasonable to question whether Hatfield's definitions can effectively target those activities which are introspective, or if they are too permissive and encompass a range of activities ordinarily thought of as just perceptual and not reliant on introspection. First, though, there is an important exegetical question over how to understand Hatfield's claim that introspection is a source of evidence in experiments such as the metameric matching one.

One possible reading, which I reject, is that Hatfield just points to the fact that psychophysics, unlike behaviourist psychology assumes and moreover requires that experimental subjects have conscious perceptual experiences[2]. For the mission of psychophysics, an experimental approach to the mind inaugurated by Fechner (1860), is to chart and measure the physical energies needed to elicit specific conscious perceptual states. But I strongly doubt that Hatfield intends to characterise methods in psychology as introspective purely in terms of a contrast with the behaviourist research program. In fact, Hatfield is in agreement with Danziger (1980) that our understanding of introspection in psychology has been distorted by the behaviourist reaction to Titchener's analytical school.

What is more, the mere having of conscious states is a different thing from the possession of some ability to report reliably on the nature of those states. It is the latter capacity that is typically identified with introspection. For example,

---

psychology more generally has become somewhat blurry in recent years, with many experiments that are classified as psychophysical dealing with complex perceptual states, not just simple sensation.

[2] Not to say that the behaviourist psychologists all assumed that human beings were unconscious zombies, but that their experimental methods were indifferent to the presence or absence of consciousness.

Draft. Please do not quote without permission. Comments welcome.

Schwitzgebel's (2011) sceptical case against introspection does not target the idea that we have conscious states, or that those states are important to our mental economy, but is concerned to argue that to a greater extent than we care to admit the contents of those states are indeterminate or unknown to us. My interpretation of Hatfield's notion of introspection will hinge on this point. I understand his claim that contemporary perceptual psychology relies on introspective evidence to be the claim that psychologists exploit their subjects' introspective ability in order to glean information about the human perceptual system, and furthermore it is a presupposition of this experimental practice that subjects are competent introspectors in the sense that they are capable of giving verbal or motor responses which reliably indicate the presence or absence of particular features of their conscious experiences. For example, a psychophysical experiment which measures the absolute detection threshold for a dim spot of light is said to be reliant on the subject's capacity to introspect in the sense that her subjective awareness of the spot is a crucial data point that the experimenter has access to because of the subject's capacity to introspect. And thus the experimenter must assume that the subject can faithfully indicate those times that the spot enters her conscious field of view.

Yet a problem with this account is that it is not clear how it can be employed to distinguish introspection from ordinary perception, for doesn't the subject's activity in the detection task just boil down to her *looking* for a dim spot of light? This worry could prompt us to take Hatfield as endorsing a more restrictive definition. For Hatfield also suggests that what characterises introspection over ordinary perception is that one attends to one's experience of an object, not just the object itself (p. 279, "immediate attention to… phenomenal experience"). This makes introspection importantly different from perception because as many would have it, perception is generally "transparent" and our perceptual encounter with the world is not interrupted with moments of attention to experience itself. The difficulty with this reading is that it then becomes unclear how the more restrictive definition of introspection could apply to many of the psychophysical tasks that Hatfield wants it to apply to, such as stimulus detection and the metameric matching experiment. Subjects perform such tasks by directing their attention to external stimuli, namely the coloured lights, and need not attend to their own phenomenal experience, *qua* experience. Nor do they need to consider their experience in a more fine grained or detailed way than in ordinary perception.

In short , the problem is that while Hatfield's restrictive definition has the virtue allowing one to demarcate introspection from perception, it cannot reasonably be applied to the range of psychophysical tasks that Hatfield claims it does. And furthermore a case could be made that it should not apply to any perceptual experiment, since these generally involve attention to external objects, not attention to phenomenal experience itself. Yet the more liberal definition makes all perceptual activity concurrently introspective in a somewhat trivial sense.

Draft. Please do not quote without permission. Comments welcome.

It strikes me that a different approach to defining introspection -- in the context of psychophysics -- is needed, one that does not characterise introspection in terms of an object of attention or focus of awareness. In this paper I propose that the tasks that should be said to involve introspection are the *ones which rely on experimental subjects' capacity to analyse and compare sensory experiences that bear non-obvious relationships of similarity and difference to each other*. Thus on my account introspection can be part of the process of perceiving and attending to an external object, and need not be overtly directed at phenomenal experience. The subject may interpret her task to be simply that of attending to the external stimulus, but she can be reporting on aspects of her phenomenal experience nonetheless. It is also a feature of my view that the extent to which tasks rely on introspection is a matter of degree. In the next section I give a set of examples of common psychophysical tasks that are either "introspection-heavy" or "minimally-introspective". In the third section I describe how the cluster of introspection-heavy tasks – though not described in this way by scientists themselves – has commonly attracted suspicion from psychophysicists as being less likely to produce data that is "objective" and informative about neural mechanisms. I ask whether this is mere coincidence, or if the methodological norms of psychophysics reflect a certain wariness towards introspection.

2. Introspection in Psychophysics as Controlled Comparison

2.1 Examples Of Introspectively Demanding and Undemanding Psychophysical Tasks.

The *metameric match paradigm*, illustrated in figure 1 has been used to diagnose specific types of colour vision deficiency since the late 19th century. Differences in the number of retinal cone types an individual has, and the spectral sensitivities of those cone classes, lead to measurable differences in the proportion of red to green in a composite light that he or she judges to look identical to a yellow monochromatic standard. Note that in this task the only perceptual judgment that the subject need make is over whether the composite and monochromatic light are visually indistinguishable. If the lights are presented as abutting (as in fig. 1) then the subject simply has to judge whether or not the colour field is homogeneous. No attention to the specific qualities of the perceived colour is required.

Draft. Please do not quote without permission. Comments welcome.

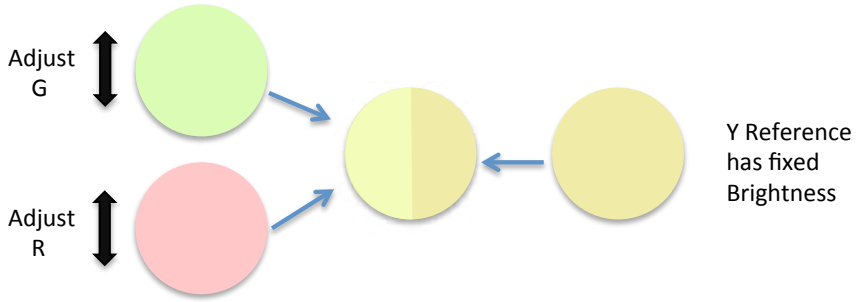


Figure 2
An Asymmetric Match Experiment. The subject is asked to adjust the proportions of R and G monochromatic lights so that the yellows match in hue. The intensity of the Y light is fixed, so the yellows cannot be matched for brightness and are therefore distinguishable even when hue is judged to be equivalent.

Contrast this task with an *asymmetric match paradigm* (fig. 2). In this case the two central stimuli are not matched for luminosity but the subject must say whether or not they match in hue regardless of their visible difference in brightness. This requires that the subject analyse her experience of the two colours in terms of separate dimensions of hue and brightness, and make a judgment as to the identity of just one of these dimensions, disregarding the difference in the other. Thus the subject must make a series of comparisons between pairs of stimuli in order to find the pair that holds the unique but non-obvious relationship of sameness of hue. This relationship is non-obvious in that it is not marked by a simple defining characteristic like a homogenous spatial profile.

It should be fairly intuitive that this task is "introspection-heavy" in a way that the metameric matching task is not. The contrast between these two tasks points us to the central intuition behind my new characterisation of introspection. The idea is that the metameric matching task is "minimally-introspective" because it can be performed without any careful comparison of the phenomenal qualities one experiences on presentation of the two stimuli. The metameric paradigm relies on introspection only in the minimal sense that it assumes the subject can know and reliably report when her conscious visual field is homogeneous with respect to colour[3]. The asymmetric matching task, on the other hand, is "introspection-heavy" because it does require this careful comparison of sensory experiences that bear non-obvious relationships of similarity and difference to each other.

Asymmetric matching paradigms have been used to study achromatic perception of lightness and darkness (fig. 3a , see Gilchrist 2004) and to study colour constancy. Figure 3b gives an example of an asymmetric task in which the observer views a

[3] i.e. relies on introspection defined in the first, permissive sense. To reiterate the discussion of section 1, the problem with the minimal notion of introspection is that it cannot distinguish introspection from ordinary perception.

Draft. Please do not quote without permission. Comments welcome.

scene under two different lighting conditions. She is instructed to adjust the colour of the central patch in one image until it looks as if made from the same paper as the central patch in the other (Foster, 2011). Importantly, even when the patches are matched there will still be a visible difference in colour between them, and the experiment relies on the subject having a clear sense of what sameness of material would look like in spite of these differences. Again, the task is "introspection-heavy" in comparison to a task in which the subject just has to report on the absolute identity or distinguishability of two stimuli. In particular, it relies on the subject's ability to make a "judgment call" on the one best match, given a range of close contenders which vary along a number of different dimensions. I describe the introspection-heavy tasks as requiring *controlled* comparison because the demand placed on the subject is to perform some kind of analysis and comparison, but within parameters that are pre-specified by the experimenter.

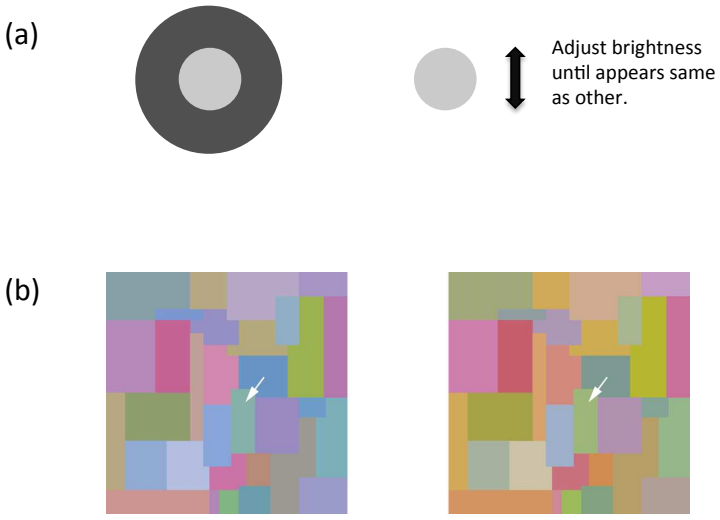
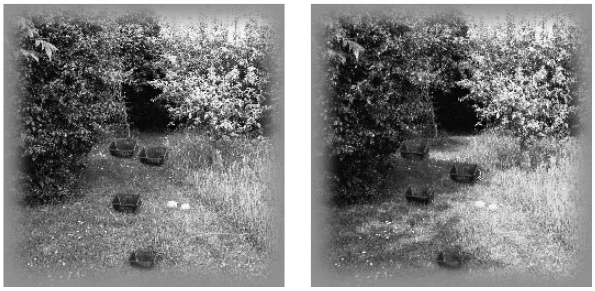


Figure 3
(a) *Achromatic asymmetric match experiment where black annulus influences perceived brightness of one of the circles. Subject is asked to determine point of subjective equality of the brightness of the two circles.*
(b) *Asymmetric colour constancy experiment. Subject is asked to adjust the colour of one of the patches (marked with arrow) until it looks as if it is made from the same paper as the other. (From Foster 2011, permission needed.)*

Another kind of paradigm that intuitively fits the idea of controlled comparison is a rating scale task. In a series of experiments published recently (To et al 2008, 2010, Tolhurst et al 2010) subjects were presented with nearly 300 pairs of photographs – an original and a modified version – and were asked to rate how similar the pairs were on a scale from 0 (completely identical) to any arbitrarily high value (see fig.

4a). In one of these publications, Tolhurst and colleagues (2010) also present results of a simple two-alternative-forced-choice (2-AFC) contrast discrimination experiment in which subjects just had to report which of a pair of otherwise identical photographs contained a small, high contrast central patch (see fig. 4b). They then apply their model of contrast discrimination to the rating scale data. The rating scale task falls under my introspection-heavy category, while the contrast discrimination task is minimally-introspective. In the former, the subject must make a judgement as to the relative similarity of a large number of pairs of stimuli, that differ in different ways, whereas in the latter task she detects the presence or absence of a high-contrast patch in a rather automatic fashion. Figure 4 illustrates how similar stimuli can be used in these two very different experiments, so it is not complexity of stimulus *per se* that determines how introspectively demanding the task is. Rather, the determining factor is the nature of the response that the subject must make to the stimulus. That is, whether the response is simply choice between saying the high contrast patch appeared first or second out of two stimuli, or if it calls for a more careful examination of the perceived properties of the stimuli.

(a)



(b)

Figure 4
Example of stimuli used by Tolhurst et al. 2010 (a) *Rating scale task. For each of 294 image pairs, subjects were asked to rate how similar or different they appeared on a numerical scale of their own devising. (permission needed)*

(b) *2-AFC contrast discrimination. Subjects had to report if the high contrast central patch appeared in the first or second stimulus.*

Before moving on, I would like to emphasise that my two categories are intended to reflect a qualitative difference in how introspectively demanding these tasks are, and that I will say nothing in this paper about how to quantify this difference, and how it is that introspective demands admit of degree. For example, the question of whether or not metameric matching is even less introspectively demanding than contrast discrimination will be left unanswered. It seems plausible that introspective demands, like attentional demands, come in degrees but I offer no suggestions of how one might measure this. It is also plausible that there will be some tasks that occupy middle ground between my categories and are hard to classify either way. I will not deal with such cases here. My aim in presenting a set of tasks that are intuitively more reliant on introspection than the others has been to highlight one way that introspection may be said to play a role in perceptual psychology, and to this end I have focused on the most clear cut cases.

2.2 Other Classifications of Psychophysical Tasks

One of the attractive things about psychophysics as a subject for philosophy of science is the fact that throughout its short history methodological questions about the best way to measure sensory responses have been debated in a perspicacious way by leading protagonists. Moreover, such controversies still resonate in the living memory of the discipline, and are recounted even in the most recent textbooks. One way in which methodological debates commonly unfold is with a distinction first being drawn between two broad classes of psychophysical techniques, and the relative merits of the two classes are subsequently discussed.

In their textbook Kingdom and Prins (2009) devote a chapter to the "dichotomies" that have been most significant to psychophysicists past and present. The first of these, Brindley's (1960, 1970) distinction between Class A and Class B observations is particularly relevant to my account of introspection. Brindley characterised Class A observations as any tasks in which the observer just had to report on the absolute similarity or dissimilarity in the appearance of a pair of stimuli. For example, the measurement of the detection threshold for a spot of light is Class A because the subject need only indicate whether the trial in which the spot is present is distinguishable or not from the reference stimulus in which the spot is absent. Likewise, the measurement of the discrimination threshold for the brightness of the spot is also Class A, as it just requires the subject to report if the trial in which the luminosity of the spot is increased looks different from the trial in which the luminosity remained at baseline. In contrast, Brindley (1970:133) categorised as Class B, "[a]ny observation that cannot be expressed as the identity or non-identity of two sensations…"; for example, "all those [observations] in which the subject must describe the quality or intensity of his sensations, or abstract from two different sensations some aspect in which they are alike."

Draft. Please do not quote without permission. Comments welcome.

Brindley's description of Class B observations is interchangeable with my characterisation of introspection-heavy tasks. Indeed, the tasks which I presented as examples of my minimally-introspective category – metameric matching and contrast discrimination – are Class A, whereas all kinds of asymmetric matching and rating scale tasks are Class B. In essence, both of these categorisation schemes can be understood as drawing a distinction between tasks in which the experimental subject is treated somewhat like a thoughtless measuring instrument, and methods that rely on the subject's status as a critical being who can attend to and reflect on her own conscious states. The point is not that the A/minimally-introspective Class treats the subject as if unconscious, or that it requires the subject to have sensory capacities but not cognitive ones. Rather, it is that the A/minimally-introspective Class makes no demands on any capacity for reflection on and comparison of occurrent sensory states, whereas tasks in the B/introspection-heavy Class do[4].

To illustrate this, imagine a machine that can read off the conscious sensory states of a subject performing a contrast discrimination task. In order to predict the subject's responses to any trial, all the machine must do is to assign a number to the intensities of the subject's experience of contrast for the central regions of the two different stimuli. If they have the same values the machine predicts the answer is 'same', and if they differ the machine predicts an answer of 'different'. Once the non-trivial problem of reading off individuals' phenomenal states is solved, the rest is uncomplicated! If an equivalent machine were to be built for the rating scale task, the blueprint could not be so simple. There is no one quality of the subject's conscious experience of the stimuli that the machine could measure and use to predict the response. Rather, the machine would have to rely on some complicated model of how various differences in the experienced qualities of the images are weighted against each other to give and impression of greater or lesser degrees of similarity[5]. In other words, a model of introspective comparison and not a simple measurement algorithm.

The mind-reading machine thought experiment again confronts us with the fact that the distinction being drawn is not between tasks that are in no way introspective and those that completely are. Rather, it is about the extent to which these tasks call upon some putative introspective capacity. For the first machine, dealing with the

[4] In support of this idea that the key distinction in play here is between subject-as-measuring-instrument and subject-as-reflective-being, it is worth noting that Brindley's one example of a psychophysical document explicitly hostile to Class B observations is the 1943 Optical Society of America (OSA) report that, as Stevens (1951:31) relates, "reduces psychophysics to the employment of a human observer as a null instrument under a set of strictly specified conditions" And Brindley's one example of a psychophysicist liberal with regards to Class B is Stevens (1951), who explicitly rejects the OSA definition as too narrow and restrictive (and cf. Helson 1949).

[5] Interestingly, however, Tolhurst et al. (2010) can predict trends in the similarity rating data to a fair degree of accuracy with a model of entirely unconscious neuronal response functions. The fact that there is "machine" that can predict responses to the contrast discrimination and rating scale experiments, without peering into the conscious states of subjects should not detract from the fact that any hypothetical machine attempting to examine conscious states in order to predict responses would have a to treat the two experiments differently.

contrast discrimination experiment, can still peer into the conscious states of observers and this captures some minimal notion of introspective activity. Yet the second machine, dealing with the rating scale experiment, needs not only to determine what the subject experiences, but also to determine what the subject makes of her experience, what is more and less salient about the different qualities presented in her visual phenomenology. This is an introspective undertaking of a weightier kind.

It is hard to say how influential Brindley's distinction has been. It came under immediate criticism from Boynton and Onley (1962) but was clearly accepted in some form by Marks (1978) and Teller (1984), and is discussed at length in Gescheider's (1997) psychophysics textbooks. Kingdom and Prins (2009, p.18) choose not to employ it as an overarching basis for classifying psychophysical experiments because of the problem that certain tasks cannot be classified as either A or B.

Kingdom and Prins' preferred distinction is between tasks that measure performance and those that measure appearance, which they characterise in the following way:

> "If the measurement can be meaningfully considered to be better under one condition than under another, then it is a performance measure, if not it is an appearance measure." (p.22)

Performance tasks are any ones designed to chart perceptual "limits" (e.g. contrast discrimination, detection of a spot of light against a differently coloured background). An example of an appearance task is an experiment comparing the strength of the Müller-Lyer illusion with fin angles of 45° and 60°. Even if the length of the central bars appears to be more different when the fins are 45°, there is no sense in which the subject is "better" at the task in that condition. So this Class B observation can also be said to be an appearance measure. Thus there is an overlap with my distinction: appearance tasks tend to be introspection-heavy, and performance tasks tend to be minimally-introspective. But it is not as well matched as is the case with Class A vs. B. In particular, the metameric match task that I classify as minimally-introspective turns out to be an appearance measure.[6]

3. Not All Psychophysical Methods Were Created Equal

All I have argued so far is that there is an intuitive way of differentiating psychophysical tasks that are more reliant on introspection from those that are not, and that my categorisations turn out to be roughly co-extensional with categorisations of tasks developed within the psychophysical tradition. The

[6] A related dichotomy is Sperling's (et al. 1990) Type 1 vs. Type 2 distinction. In Type 1 experiments the subject's response maybe either correct or incorrect with respect to some physical dimension of the stimulus (e.g. for either is more oblique than line 2). For Type 2 the experimenter is cannot classify responses as correct or incorrect. Note again that the metameric match turns out to be Type 2, even though Class A/minimally-introspective.

question now is what to make of this finding. Is it just a coincidence that the distinctions coincide? It should come as no surprise to the reader that my next point will be that the categories that line up on the introspection-heavy side have tended to meet with more diffidence and suspicion from psychophysicists than those on the other side.

Brindley (1970) presents type A observations as especially informative about the physiological mechanisms underlying perception because they can be used to test "psycho-physical linking hypotheses", that two stimuli (e.g. yellow monochromatic light, and a certain mixture of red and green lights) will produce the same neural activity and hence the same sensation.

On the relative status of the two classes he writes that,

> "The use of Class A observations as a basis for analysing the function of the eye and visual pathway is not controversial; every writer on vision admits, at least by implication, that they can be legitimately used. On the use of the kinds of observation here called Class B, there have been differences of opinion … The conservative opinion, in its most extreme form, is that only Class A observations are of any value, and in a discussion of visual mechanisms all Class B observations may be entirely disregarded." (1970, p. 134)

Brindley himself takes this view to be too narrow, but is critical of Stevens' (1951) "extreme liberal opinion" for failing to make the distinction. Later in the book, when discussing Hering's opponent theory of colour he writes as if it is still moot whether the kinds of phenomenological reports presented by Hering in support of his theory can actually be taken as evidence for a kind of colour mechanism (p.208).

One might think that this is all besides the point in a discussion about introspection because the reason why the value of Class B observations was held in question was not because they are introspection-heavy, but because their failure to underwrite psychophysical bridge principles. But I do not think that this problem is so disconnected to from the issue of introspection. For if Class B tasks were to be granted some supporting assumptions, like the ones offered for Class A, then one could equally say that they are informative of underlying neural mechanisms. For example, in the case of the asymmetric hue matching experiment, why not assume that when the hue sensation for each stimulus is equal, that is evidence that there is a neural pathway somewhere between the photoreceptors and the cortex that conveys the same message in both cases? This would be a special case of the assumption made in support of inferences from Class A observations that, "whenever two stimuli cause physically indistinguishable signals to be sent from the sense organs to the brain, the sensations produced by these stimuli, as reported by the subject in words, symbols or actions, must also be indistinguishable" (Brindley 1970, p.133).

Yet, Class B observations *are* treated differently. The reason for this difference is likely because Brindley and other theorists (e.g Marks 1978) have been wary of attributing to subjects the kind of introspective powers that would be needed to

analyse hue separately from all other sensory qualities, and determine exactly the point of equivalence of hue. In other words, if these theorists had shared Titchener's faith in the analytical acumen of introspection, they would have had no reason to treat Class B observations differently from Class A.

This pattern of unequal treatment can be seen not just in the discussion of Class A and B observations, but also with respect to the other dichotomies discussed by Kingdom and Prins. They note that it is fairly common for psychophysicists to refer to some tasks as more "objective" or "subjective" than others, with all the value-laden connotations of these terms. Kingdom and Prins explain this usage in the following way:

> "All psychophysical experiments are in a trivial sense subjective, because they measure what is going on inside the head, and if this is the intended meaning of the term then the distinction is redundant[7]. The dichotomy is more often invoked, however, to differentiate between different types of psychophysical procedure. The distinction has been used variously to characterize Class A versus Class B observations, tasks for which there is versus tasks for which there is not a correct and an incorrect response[8], forced-choice versus non-forced-choice procedures, and criterion-dependent versus criterion-free procedures." (p.18-19)

The notion of subjectivity at play here is encapsulated in the idea that experiments are subjective if they are introspection-heavy. For all the tasks on the wrong side of the subjective-objective tracks are ones which rely on the subject's judgments concerning the appearance of the stimuli, involving complex comparisons which cannot be independently verified by examining the physical properties of the stimuli themselves.

To conclude, there is a sense in which the title of this paper is misleading. I have not showed that the psychophysicists have avoided using experimental methods more reliant on introspection, or that the use of such methods has always been questioned. Indeed, when Kingdom and Prins write that, "Both performance-based and appearance-based experiments are important to our understanding of vision. Measures from both types of experiments are probably necessary to fully characterize the system" (p.26), they are articulating a methodological pluralism that many psychophysicists would endorse. However, the crucial point is that the methods on the wrong side of the divide, those more reliant on introspection, continue to need their advocates, whereas those on the other have been accepted without question. This is an indication of the contested status of introspection within the psychophysics tradition.

---

[7] Cf. the worry discussed above that all psychophysical experiments rely on introspection in a trivial or "minimal" way, hence the distinction between introspection and perception is made redundant.
[8] I.e. performance vs. appearance or Sperling's Type 1 vs. Type 2.

Draft. Please do not quote without permission. Comments welcome.

References

Boynton, R. M. and J. W. Onley (1962). "A critique of the special status assigned by Brindley to 'Psychophysical Linking Hypotheses' of 'Class A'." Vision Research 2: 383-390.

Brindley, G. S. (1960, 2nd edition 1970). Physiology of the Retina and the Visual Pathway. London, Edward Arnold.

Danziger, K. (1980). "The History of Introspection Reconsidered." Journal for the History of the Behavioural Sciences 16: 241-262.

Fechner, G. (1860/1966). Elements of Psychophysics Holt, Rinehard and Winston.

Foster, D. H. (2011). "Color Constancy." Vision Research 51: 674-700.

Gescheider, G. A. (1997). Psychophysics: The fundamentals. Mahwah NJ, Lawrence Erlbaum.

Gilchrist, A. (2004). Seeing Black and White. Oxford, Oxford University Press.

Hatfield, G. (2005). Introspective Evidence in Psychology. Scientific Evidence: Philosophical theories and applications. P. Achinstein. Baltimore, Johns Hopkins University Press.

Helson, H. (1949). "Review of 'Introduction to Color'." Psychological Bulletin 46(2): 166-169.

Kingdom, F. A. A. and N. Prins (2009). Psychophysics: A practical introduction. Amsterdam, Elsevier Academic Press.

Marks, L. E. (1978). The Unity of the Senses. New York, Academic Press

Schwitzgebel, E. (2011). Perplexities of Consciousness. Cambridge MA, MIT Press.

Sperling, G., B. A. Dosher, et al. (1990). "How to study the kinetic depth experimentally " Journal of Experimental Psychology: Human Perception and Performance 16: 445-450.

Stevens, S. S. (1951). Handbook of Experimental Psychology. London, Chapman & Hall

Teller, D. Y. (1984). "Linking Propositions." Vision Research 24(10): 1233-1246.

To, M. P. S., P. G. Lovell, et al. (2008). "Summation of perceptual cues in natural visual scenes." Proc. Royal. Soc. Lond. B. Biol. Sci 275: 2299-2308.

To, M. P. S., P. G. Lovell, et al. (2010). "Perception of suprathreshold naturalistic changes in colored natural images." J. Vision 10: 1-22.

Tolhurst, D. J., M. P. S. To, et al. (2010). "Magnitude of Perceived Change in Natural Images May be Linearly Proportional to Differences in Neuronal Firing Rates." Seeing and Perceiving 23: 349-372. Reprinted in J.A. Soloman (ed) 2011, *Fechner's Legacy in Psychology*. Boston: Brill

Individual Based Models in Ecology:
An Evaluation, or How Not to Ruin a Good Thing

Joan Roughgarden
Professor Emerita, Dept. of Biology, Stanford University
Adjunct Professor, Hawai`i Institute of Marine Biology, U. of Hawai`i

For: Philosophy of Science Association
Biennial Meeting, San Diego, 2012

Manuscript: November 9, 2012

Background: What are now increasingly called individual based models (IBMs) have been used in ecology since the 1970s when theoretical ecology began in earnest. The best known examples from that time include the forest computer simulation model (named JABOWA) of Daniel Botkin (Botkin et al 1972) and the computer simulation model of Donald DeAngelis (DeAngelis et al 1991) for a freshwater fish cohort. These were identified with the systems ecology school of theoretical ecology and the approach was anticipated to offer a unifying theoretical framework for ecology (Huston 1988), a goal whose possibility is still being debated (Roughgarden 2009, Vellend 2010). Since then hundreds of IBMs have been published in ecology. Moreover, IBMs are being actively developed in other disciplines, especially the social sciences, and dozens of software environments have been created to facilitate IBM research (Allen 2010, Borrill and Tesfatsion 2012). This talk reviews progress for IBMs in ecology, details several remaining difficulties, and suggests clarification where needed.

Provisional Definition: For now, an IBM is provisionally considered to be a computer simulation in discrete time steps for the creation, disappearance and movement of a finite collection of discrete interacting entities. The germination, growth and death of a collection of individual trees on a plot of ground, or the birth, growth and reproduction of a collection of individual fish in a pond are the classic examples.

Challenges Met: Grimm and Railsback (2005) detail seven “challenges” that IBMs have faced in ecology: long time needed to develop the model, difficulty in analyzing results, lack of common language to communicate model and results, requirement for too much data, uncertainty and error propagation, lack of generality, lack of standards. Ecological IBM modelers have faced these challenges head on. They have collectively proposed and implemented a protocol (called the “ODD protocol”) for how a model is to be specified (Grimm et al 2006, 2010), and they have coalesced around a freely downloadable programming platform, NetLogo (Wilensky 1999, 2013), as a standard for developing and executing IBMs (Lytinen and Railsback 201, Railsback and Grimm 2011). Moreover, NetLogo can be embedded within Mathematica (Wolfram Research) thereby endowing the IBM modeling module with the statistical and analytical tools of Mathematica’s powerful industry-standard mathematical programming environment. The steps by the ecological IBM modeling peer group go a long way toward resolving many,

but not all, of the reservations that have dogged IBMs since their inception. Here are some remaining problems.

Exclusionary Definitions: Despite their progress, ecological IBM modelers have also taken decisions that seem counter productive. They employ an unnecessarily exclusive definition of what counts as an IBM. Grimm and Railsback (2005), following Uchmanski and Grimm (1997), stipulate that to be considered an IBM in ecology, the model must satisfy four criteria:

1. Detail about each individual's life cycle must be present in the model, including the growth and development of each individual as it ages.
2. The dynamics of resources used by individuals must be explicitly represented - a "carrying capacity" cannot be used because it is supposedly a population-level concept and that cannot be known to an individual.
3. Integers and not real numbers must used to represent the size of a population-the model must feature discrete events and not refer to rates.
4. Variability must be allowed and must exist among individuals of the same age - environmental phenotypic variation, not heritable genetic variation, in as much as Grimm and Railsback (2005) consider evolutionary ecology as beyond the scope ecological IBMs.

Inconsistent Definitions: Grimm and Railsback (2005) acknowledge that these criteria rule out many models as IBMs. Notable among the excluded models are “predator-prey systems with individuals as discrete units with local interactions but no life cycles or variability among individuals”. However, this criterion conflicts with standard practice in the wider IBM community. Wolfram’s Mathematica website has a demonstration by Sayama (no date given) of a “real-time agent-based simulation of a predator-prey ecosystem” wherein rabbits run around in a square area and are chased by foxes. Castiglione (2006), in the Scholarpedia peer-reviewed open-access encyclopedia entry about agent-based modeling, also features a direct comparison of an an individual-based fox-rabbit model compared with the venerable Volterra predator/prey model that is formulated as pair of coupled non-linear differential equations. The fox-rabbit models proposed as examples of IBMs would nonetheless be ruled out as ecological IBMs by Grimm and Railsback even though they are offered precisely as illustrations of IBMs in the wider IBM literature.

Why So Restrictive? In acknowledging that their definition is restrictive, Grimm and Railsback (2005) refer to models that seem in some respect to be IBMish but are not true IBMs, as “individual-oriented”. Why do Grimm and Railsback care so much about retaining their exclusionary definition? Because they are committed to the ideal that “IBMs can lead to a fundamentally new view of ecological systems and processes”. They write that unlike true IBMs, “individually-oriented models do not allow us to fully trace the systems properties back to the behavior of the individual animals". The ecological IBM modelers regularly disparage the “classical framework” for describing ecological systems as “relatively simple and characterized by system-level state variables”, vs “the IBM view that ecological processes and systems emerge from the

traits of adaptive individuals", and they view their exclusionary definition of an IBM as necessary to accomplish this aim. Let us consider then whether the restrictions are in fact necessary to attaining a "fundamentally new view of ecological systems."

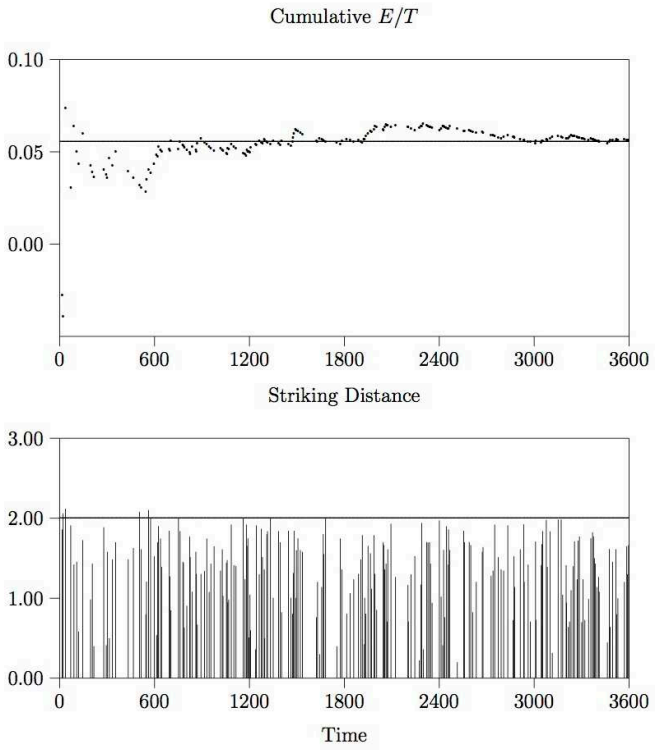


Figure 1.15: Top: Net energy gained per elapsed time in joules/s during one h of foraging, as noted after each prey capture. Expected yield from optimal home range shown as horizontal line. Bottom: Distance in m of successful strikes as function of time in s. Optimal cutoff radius shown as horizontal line. Insect length is 2.5 mm, prey flightiness coefficient is 1, abundance is 480 insects per m^2 per 12 h, lizard SVL is 45 mm, and lizard's memory extends to beginning of foraging period.

"Individually Oriented" Models Are Sufficient: I now review two examples of models that are IBMish but do not satisfy Grimm and Railsback's (2005) criteria, and show that these do represent a fundamentally new approach to formulating ecological models.

(1) Optimal Size of an Optimal Forager: In 1995 I published a model for how a lizard could learn to forage optimally (Roughgarden, 1995). The model predicted the "optimal cutoff distance" such that all prey closer than this distance are taken and all prey beyond this distance are ignored. The optimal distance is that which maximizes the lizard's rate of energy capture. A simple algorithm was exhibited that would allow a lizard to dynamically learn where the optimal cutoff was. The figure above illustrates the model using parameters estimated from field data for *Anolis* lizards in the Eastern Caribbean. The lower panel shows the optimal cutoff distance as a horizontal line. Prey are appearing randomly at distances from 0 to 3 m away from the lizard. Each vertical line represents a prey item that was chased and caught. Notice that vertical lines rarely cross the optimal cutoff, and those that do are principally at the beginning of the simulation when the lizard is still learning where the optimal cutoff distance is. The upper panel shows how the lizard's energy capture rate within a day approaches the optimal capture rate, shown as a horizontal line. The realized capture rate fluctuates initially reflecting the lizard making mistakes by chasing insects beyond the optimal cutoff distance or ignoring insects in front of the optimal cutoff distance. The existence and quantitative properties of the optimal cutoff distance were tested and confirmed in field studies of *Anolis* lizards on the island of Anguilla (Shafir and Roughgarden 1998).

Based on this model for the daily energy capture by a lizard, the daily growth rate of a lizard could be predicted. The next figure shows a scatter plot of lizards' daily growth increments from field data compared to that theoretically predicted assuming a lizard is an optimal forager. The open circles pertain to females and the closed circles to males. The theoretically predicted optimal growth rate is the solid curve. Notice the quantitative agreement between actual growth increments and that expected from optimal foraging theory. Females cease growing and drop off the curve when they have reached a length of about 45 mm and the males drop off the growth curve at about 60 mm in length. These sizes are typical of adults on those Eastern Caribbean islands with only one species of anole (the so-called solitary size).

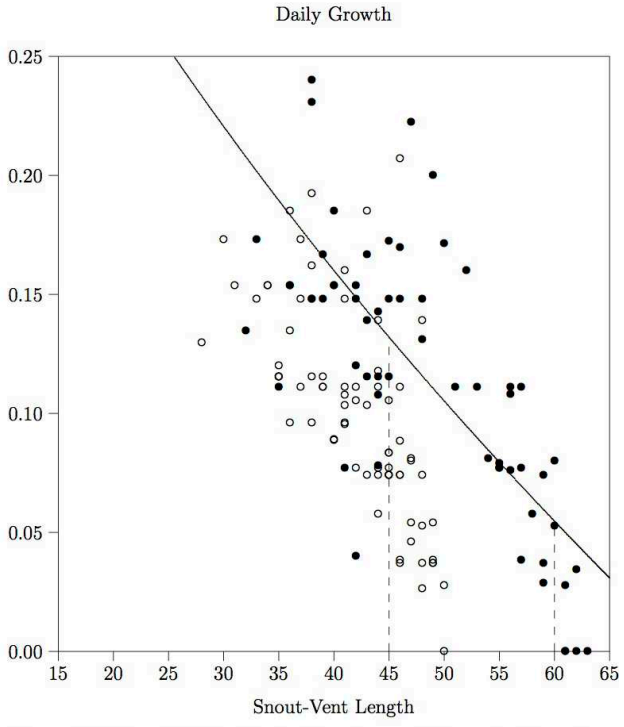


Figure 1.22: Growth of *Anolis gingivinus* in mm/d for males (•) and females (○) in St. Martin from December 1980 to April 1981 as a function of a lizard's snout-vent length in mm. Solid descending curve is growth predicted from optimal foraging theory assuming 3.5 h foraging time per d and using the optimal foraging parameters of Figure 1.16. Vertical dashed lines are conjectured switches from growth to reproduction, for females at 45 mm and males at 60 mm.

The next task is to predict why the lizards stop growing at the sizes they do in order to begin reproduction at that time. To accomplish this, the optimal growth rate curve can be integrated through time to yield a predicted curve of how the size of a lizard changes as it ages as shown in the adjacent figure.

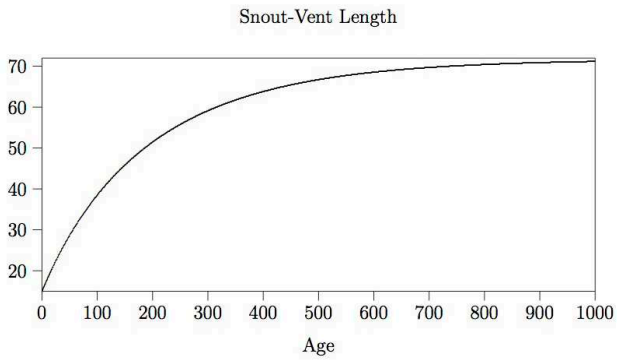
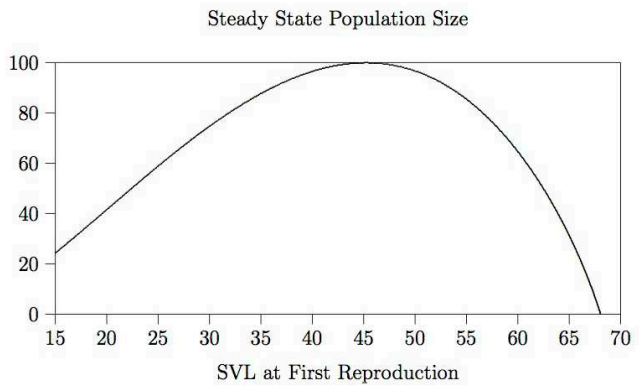


Figure 1.23: Snout-vent length in mm as a function of age in days assuming 3.5 hrs foraging hours per day, every day.

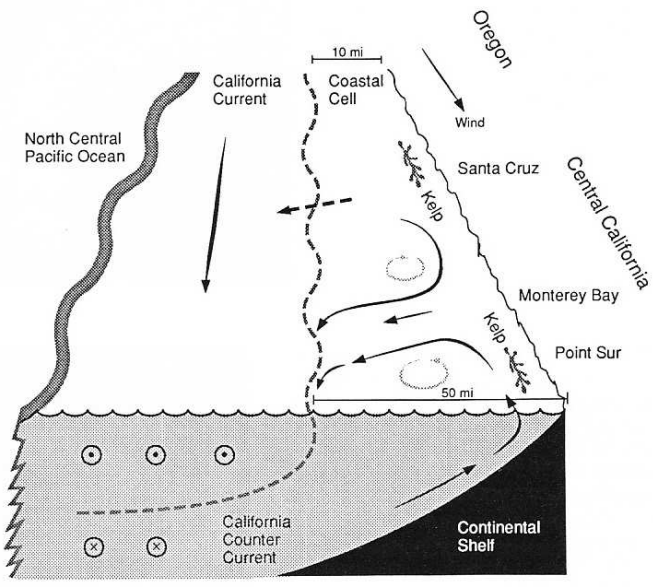
This theoretically predicted growth curve is then combined with field estimates of survivorship and with a maternity function predicted from the fecundity of an



optimally foraging female as a function of the age at which she stops growth and switches to producing eggs. Further, assuming density dependence consistent with field data showing a maximum abundance of 100 lizards per 100 sq m, the steady state abundance as a function of body size is predicted. According to density-dependent natural selection theory (*K*-selection) the body size that maximizes the steady-state abundance is the optimal body size. The figure above shows the optimal body size for females to be about 45 mm, as in fact observed. This example illustrates a complete and successful modeling protocol that begins with properties of an individual and culminates in an evolutionary prediction of the adult body size for lizards on an island in the absence of congeneric competitors.

The logic to this model is clearly bottom-up and in the spirit of deriving population-level predictions from the explicit properties of individuals. Nonetheless, this model fails every one of the four Grimm/Railsback criteria. It would be considered as “individually oriented”, although not an IBM per se.

(2) Population dynamics of barnacles on an open stretch of rocky intertidal habitat. The figure below offers a schematic diagram of the system of ocean currents off the coast of California and Oregon. Barnacles are small crustaceans whose adult phase lives attached to rocks in the zone between low and high tides. These animals release tiny shrimp-like larvae that live in the surface waters eating phytoplankton until they grow to a size large enough to attach to a rock, whereupon they metamorphose into adults. I developed a model for the population dynamics of these organisms (Roughgarden et al 1988). In the model, one equation pertains to the rate at which larvae settle out from the water onto vacant space on rocks.



Another equation pertains to the flow of larvae in the offshore currents. These two equations are coupled at the ocean-land boundary. Together they express a model for the population dynamics of barnacles. This model is formulated using a bottom-up logic based on the mechanisms for occupying space and the release of space following mortality. This model might is a "mechanism based model," or MBM, but the state variables are the number of barnacles per area of rock and the number of larvae per surface area of ocean, both of which are real numbers not restricted to integers. This model too fails to satisfy any of the Grimm/Railsback criteria, but could be considered "individually oriented" although not an IBM as such.

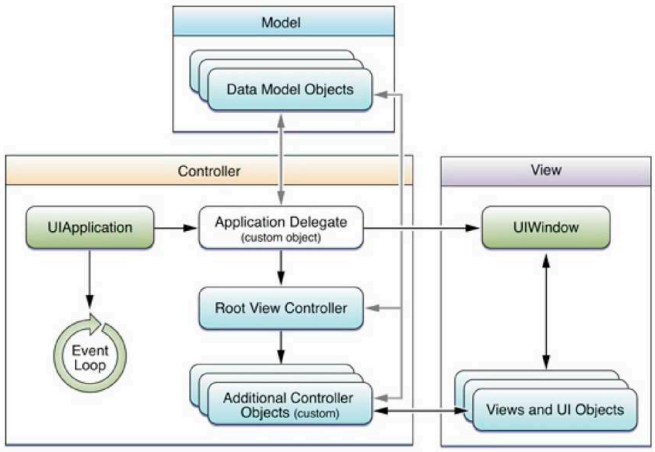
These examples show that "individually oriented" models are sufficient to achieve the goal of a "fundamentally new view of ecological systems and processes" as compared with the differential equations of classic population biology dating to the 1940's and earlier. In contrast, IBMs as defined by Grimm and Railsback seem primarily applicable only to very large organisms such as vertebrates and trees, and even then might be worthwhile only for special applications where the individuals are each specifically identified, tagged and tracked.

Individuality Confused with Agency: The difference between an individual based model and an agent based model (ABM) is confusing, with most workers considering these terms to be synonymous. For example, Castiglione (2006) writes, "An entity is an `agent' if it has some degree of autonomy, that is, if it is distinguishable from its environment by some kind of spatial, temporal, or functional attribute. That is, an agent must be identifiable. Moreover, we further require that agents must have some autonomy of action and that they must be able to engage in tasks in an environment without direct external control." Thus, identifiability and autonomy make an entity an agent in the IBM literature. So in this sense, "agent" and "individual" are roughly equivalent. Similarly, Peck (2012) writes "I follow Railsback and Grimm and make no distinction" between IBMs and ABMs. He adds that "grains of sand ... might be considered model agents ... although they do not make choices."

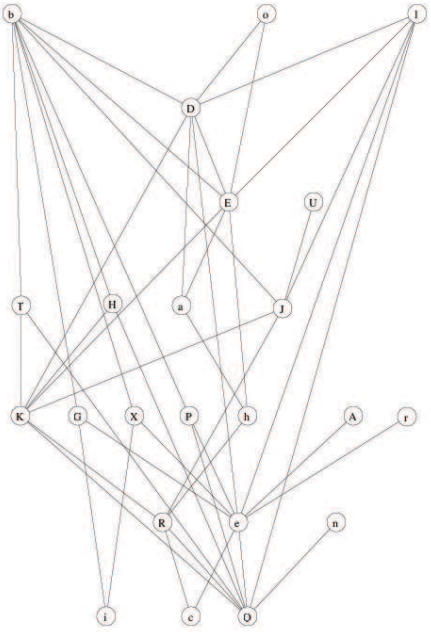
However, I think it is better to use the term "agent" more narrowly--to refer specifically to a goal-seeking individual, where the goal is to increase the individual's fitness, such as the optimally foraging lizard mentioned above. Furthermore, I require that prior to each realized action, an individual has a choice of one or more alternatives and choses the action it carries out according to the criterion that (it thinks) its fitness would thereby increase. So, to most workers an IBM and ABM are synonymous, where as in my definition, an ABM is a subset of IBM in which the individual chooses actions to pursue the goal of increasing its fitness. Choice and fitness-seeking define a biological agent.

Individuality Confused with Programming Metaphor: The definition of an IBM that most workers employ anticipates that the model will be developed using object oriented programming methods. The figure below is drawn from an Apple Computer publication about programming for the iPhone and iPad using the language Objective-C (Apple Inc. 2010). The idea, say for a hand-calculator application, is that a constellation of objects exists, such as the number and function keys together with a viewing screen as well as

some entities called a controller and a model. An event transpires such as someone pressing a key like “7”, which triggers a controller to causes a “7” to be displayed on the view. If another “7” is pressed the controller causes another “7” to be displayed, and if a “+” is pressed the controller sends the previous numbers to a model who adds them and sends the result back to the controller for display. The point here is that this programming metaphor envisions a bunch of interacting agents each with unique capabilities that collectively produce realize a function, such as a hand-calculator application, not immediately evident from inspecting the properties of the individual agents. The notion of a hand-calculator could be said to emerge from the aggregate action of the constituent components. However, what the calculator does in any instance depends on random events. The calculator just sits there endlessly, so to speak, awaiting random keys presses from a user, and then exhibiting results, all without any direction.



The object-oriented programming metaphor differs from the procedural programming metaphor, which is perhaps best envisioned with the analogy of a recipe for cooking. Indeed, the now-classic language, Pascal, is explicitly set up to enforce writing a program like writing a recipe: list the ingredients at the beginning--what the variables are and what operations are allowed on them, and then move to how the ingredients are combined to produce a chocolate cake. Procedural programming envisions a directionality, from input to output, from beginning to end.



Species coding: A. Adult spider D. *Anolis gingivinus* E. *Anolis pogus* G. Bananaquit H. Bullfinch J. Big floor insects K. Small floor insects P. Elaenia Q. Fruit and seeds R. Fungi T. Grassquit U. Gray Kingbird X. Hummingbirds a. Juvenile spider b. Kestrel c. Leaves e. Canopy insects h. Tiny floor insects i. Nectar and floral l. Pearly-Eyed Thrasher n. Scaly-Breasted Thrasher o. Nematodes r. Yellow Warbler

Both these programming metaphors are useful in ecology, but should not be confused with the issue of whether a model is formulated bottom-up (ie, “individually oriented”) vs top down. Indeed, consider the populations comprising a food web. The figure above illustrates a simplified version of a complex food web for the terrestrial community on St. Martin in the Eastern Caribbean (Roughgarden 1995). Like the hand-calculator previously mentioned, a community just sits there. Something happens to one component, like a rain squall that causes the insects on the forest floor to prosper, which in turn causes the spiders and lizards to prosper, which in turn causes the kestrel to prosper, which in turn causes increased deposition to the detritus layer and so on. The community sits there, bubbling away, without any direction--a perfect system for object-oriented programming where the populations in the community are the objects.

In contrast, a biological population is a directional entity. It grows in abundance, and adapts through evolution--a perfect system for procedural programming. It is ironic that object-oriented IBMs have been applied to population dynamics when the natural application of the approach is to communities. In any case, the value of an object-oriented programming vs a procedural programming metaphor should not be confused with the value of a bottom-up individual-oriented protocol vs a top-down protocol for model formulation

Conclusion: IBMs and ABMs originated in the 1960s when mainframe computers were first becoming available to ecological researchers. These computers provoked interest in using computer simulation for ecological modeling rather than relying on mathematical analysis. In judging the merits of model craftsmanship based on simulation vs analysis, I usually come down on the side of analysis. With simulation it may be impossible to drill down to what assumptions are responsible for conclusions, to discern the causal connections between initial conditions and results, and simulation invites unsophisticated and sloppy research together with naive hocus-pocus about the magic of emergence.

Ecological workers with IBMs and ABMs not only bear the burden of avoiding an uncritical embrace computer simulation, they risk shooting themselves in the foot. First, they propose unnecessarily restrictive definitions of what can count as an IBM, definitions that turn out to be inconsistent with usage of IBM workers in different domains. Second, they fail to distinguish between a living organism who acts through choice to increase its fitness and a dead particle. Third they confuse taking an individual organism as the conceptual starting point for ecological theorizing with the choice of programming metaphor--object oriented programming vs procedural programming. Ecological IBM and ABM workers need to clean up their act on these matters least they ruin a good thing.

Specifically, I recommend that the following definitions be adopted: (1) an IBM shall be any model for which the properties of the higher level are derived from properties at the lower level--ie, an IBM is any model formulated with bottom-up logic, any model that is “individual oriented”. (2) an ABM shall be any IBM in which the individuals at the lower

level are goal-seeking and take actions based on choices that maximize their goal. (3) Use of object-oriented programming vs procedural programming shall be considered irrelevant to the designation of a model as an IBM or ABM, and shall be undertaken according to what seems most natural to the application.

IBMs and ABMs, as distinct from computer simulation itself, offer three new conceptual advantages. First, they emphasize and implement a bottom up style of formulating ecological models--from a lower level to a higher level, eg., from an individual to social groups and thence to a population, or from organs to an organism. This perspective contrasts with traditional modeling in theoretical ecology based on the logistic and Lotka-Volterra competition and predator-prey equations. It also contrasts with the top-down approach to animal behavior required by Maynard Smith's (1982) population-genetic based solution concept of the evolutionarily-stable strategy (ESS), a approach that begins with the population's gene pool and trickles down to individual behavior.

Second, IBMs and ABMs stress an alternative programming metaphor for ecological systems--the metaphor of object-oriented programming rather than procedural programming. This metaphor seems best for modeling ecological communities where the "objects" are species united through a common food (or interaction) web, and not for modeling populations whose dynamics still seem best represented through a procedural programming metaphor that represents the directionality of population growth and natural selection.

Third, the use of ABMs strongly endorses taking the individual as the fundamental focal or "first class" object for ecology and evolution--working up from the individual to populations and communities or down from the individual to the genes within them. Resting evolutionary theory on ABMs would contrast starkly with population genetics that takes the gene as the fundamental object, and works up from there to the phenotype, the population, and beyond. The agent oriented approach in ecology contradicts the widely shared perspective in evolutionary biology that, as Dawkins (1976) articulated, ``Our genes made us. We animals exist for their preservation and are nothing more than their throwaway survival machines." Instead, according to agent-based ecology, whole individuals are the primary actors on the evolutionary stage, and the genes within them but a stage crew of temporary workers hitchhiking along for the evolutionary ride.

References

Allan, R. 2010. Survey of agent based modelling and simulation tools. Science and Technology Facilities Council (UK). Technical Report DL-TR-2010-007

Apple Inc. 2010. iOS Application Programming Guide.

Borrill, Paul L. and Leigh Tesfatsion, "Agent-Based Modeling: The Right Mathematics for the Social Sciences?" pp. 228-258 in J.B. Davis and D.W. Hands (eds.), Elgar

Companion to Recent Economic Methodology, Edward Elgar Publishers, February 2012, 560pp. ISBN-13: 9781848447547

Botkin, D. B., Janak, J. F., and Wallis. J. R.: 1972, 'Some Ecological Consequences of a Computer Model of Forest Growth, J. Ecol. 60, 849–872.

Castiglione Filippo 2006, Agent based modeling Scholarpedia, 1(10):1562 (http://www.scholarpedia.org/article/Agent_based_modeling)

Dawkins, R. 1976. The Selfish Gene. Oxford: Oxford University Press.

DeAngelis, D., Godbout, L. and Shutter, B. J. 1991, An individual-based approach to predicting density-dependent dynamics in smallmouth bass populations. Ecological Modelling 57, 91 – 115.

Grimm, Volker and Railsback, Steven F. 2005. Individual-based Modeling and Ecology. Princeton University Press.

Grimm V, Berger U, Bastiansen F, Eliassen S, Ginot V, Giske J, Goss-Custard J, Grand T, Heinz S K, Huse G, Huth A, Jepsen J U, Jørgensen C, Mooij W M, Müller B, Pe'er G, Piou C, Railsback S F, Robbins A M, Robbins M M, Rossmanith E, Rüger N, Strand E, Souissi S, Stillman R A, Vabø R, Visser U and DeAngelis D L. 2006. A standard protocol for describing individual-based and agent-based models. Ecological Modelling 198 (1-2), 115-126. [doi:10.1016/j.ecolmodel.2006.04.023]

Grimm V, Berger U, DeAngelis D L, Polhill J G, Giske J and Railsback S F. 2010. The ODD protocol: A review and first update. Ecological Modelling 221 (23), 2760-2768. [doi:10.1016/j.ecolmodel.2010.08.019]

Huston, M.A., D.L. DeAngelis, and W.M. Post. 1988. New computer models unify ecological theory. BioScience 38: 682-692

Lytinen, Steven L. and Steven F. Railsback 2011. The Evolution of Agent-based Simulation Platforms: A Review of NetLogo 5.0 and ReLogo. EMCSR (European Meetings on Cybernetics and Systems Research).

Maynard Smith, J. 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge.

Peck, S. L. 2012. Agent-based models as fictive instantiations of ecological processes. Philos Theor Biol (2012) 4:e303

Railsback, S. F., and Grimm,V. 2011. Agent-based and individual-based modeling: A practical introduction. Princeton University Press.

Railsback, Steven F., Steven L. Lytinen and Stephen K. Jackson 2006 Agent-based Simulation Platforms: Review and Development Recommendations SIMULATION 82: 609--623

Roughgarden, J., S. Gaines, and H. Possingham. 1988. Recruitment dynamics in complex life cycles. Science 241:1460-1466.

Roughgarden, J. 1995. Anolis Lizards of the Caribbean: Ecology, Evolution, and Plate Tectonics. Oxford University Press

Roughgarden, J. 2009. Is there a general theory of community ecology? Biology & Philosophy 24: 521-529.

Sayama (no date given) http://demonstrations.wolfram.com/PredatorPreyEcosystemARealTimeAgentBasedSimulation/

Shafir, S. and J. Roughgarden. 1998. Testing predictions of foraging theory for a sit-and-wait forager, Anolis gingivinus. Behavioral Ecology 9:74-84.

Uchman´ski J, Grimm V (1996) Individual-based modelling in ecology: what makes the difference? Trend Ecol Evol 11: 437–441

Vellend, M. 2010. Conceptual synthesis in community ecology. The Quarterly Review of Biology 85:183-206.

Wilensky, U. (1999). GasLab: An extensible modeling toolkit for exploring micro- and macro- views of gases. In N. Roberts, W. Feurzeig, & B. Hunter (Eds.), Computer modeling and simulation in science education (pp. 151-178). Berlin: Springer Verlag.

Wilensky, U., & Rand, W. (in press). An introduction to agent-based modeling: Modeling natural, social and engineered complex systems with NetLogo. Cambridge, MA: MIT Press.