Philosophy of Science Assoc. 23rd Biennial Mtg

San Diego, CA

Version: 30 October 2012

Table of Contents

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

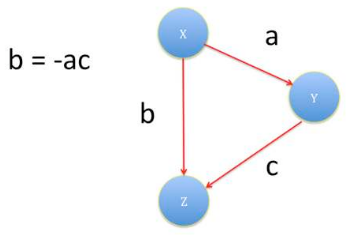
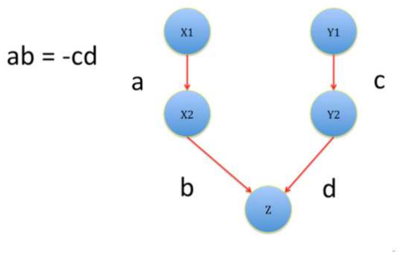


Figure 1b



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±ε_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.

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.

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

$$\begin{aligned} &\hat\rho^t_S = \textstyle\sum_i W_i(t)\; \hat\rho^0_S\; W_i^\dagger(t); \\ &\textstyle\sum_i W_i^\dagger(t) W_i(t) = I. \end{aligned} \tag{4}$$

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:

$$\rho^t_S(A) = \rho^0_S(A^t), \tag{5}$$

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

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

The probability that the state transition will be $\mathcal{T}_i$ is 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 $\mathcal{T} : \mathcal{A} \rightarrow \mathcal{B}$, there is a dual map $\mathcal{T}^* : \mathcal{A}^* \rightarrow \mathcal{B}^*$, 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 $\mathcal{T}^* : \mathcal{K}(\mathcal{A}) \rightarrow \mathcal{K}(\mathcal{B})$, there are operators $W_i : \mathcal{H}_\mathcal{A} \rightarrow \mathcal{H}_\mathcal{B}$ such that

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

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:

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

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

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

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,

$\rho_S = \text{tr}_R \rho_{SR}$ 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.

The Value of Cognitive Values

Heather Douglas

To be presented at PSA 2012 and published in the proceedings

Word Count: 4914

<u>*Abstract (100 words):*</u>
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.

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 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).

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 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

$$\frac{dN}{dt} = rN(1-\frac{N}{K}) \qquad (1)$$

(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*).

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.

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.

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-19th/early-20th 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 1 Simple stem cell model: (a) single cell, (b) cell population.

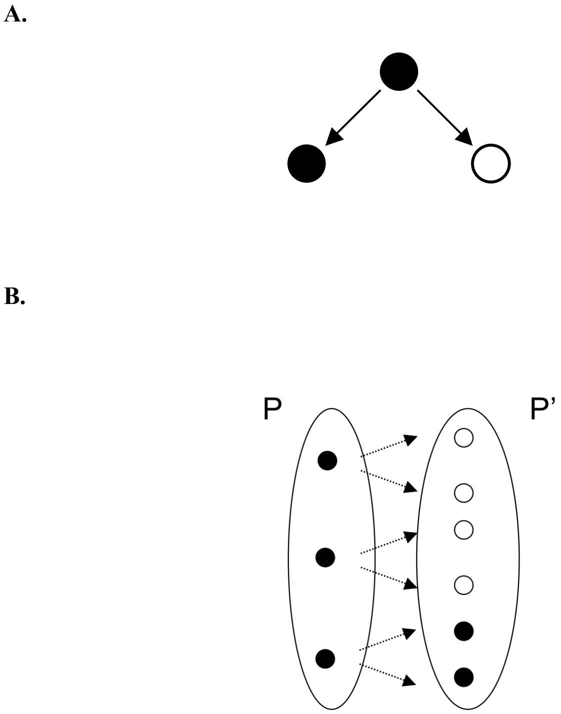


Figure 2 The stem cell concept: (a) self-renewal, (b) differentiation, (c) both. Arrows represent cell reproductive processes, variables represent key parameters (see text).

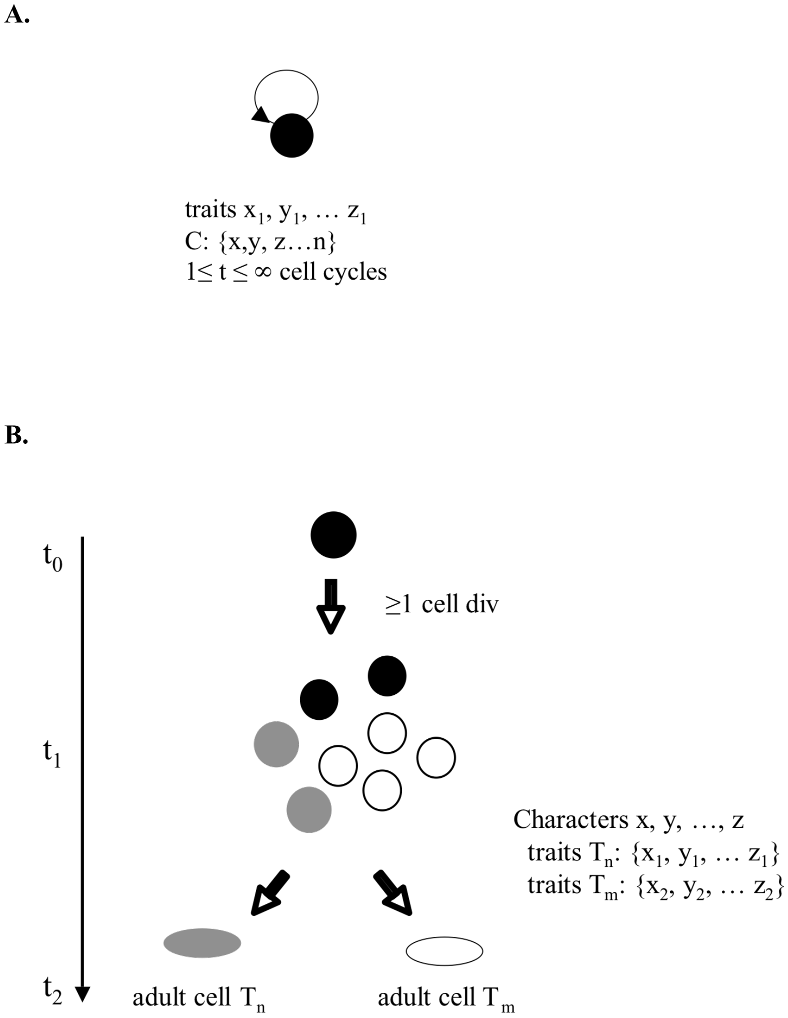

A.
traits x_1, y_1, … z_1
C: {x,y, z…n}
$1 \leq t \leq \infty$ cell cycles
B.
t_0
≥ 1 cell div
t_1
Characters x, y, …, z
traits T_n: $\{x_1, y_1, \ldots z_1\}$
traits T_m: $\{x_2, y_2, \ldots z_2\}$
t_2
adult cell T_n
adult cell T_m

Figure 2, cont.

C.

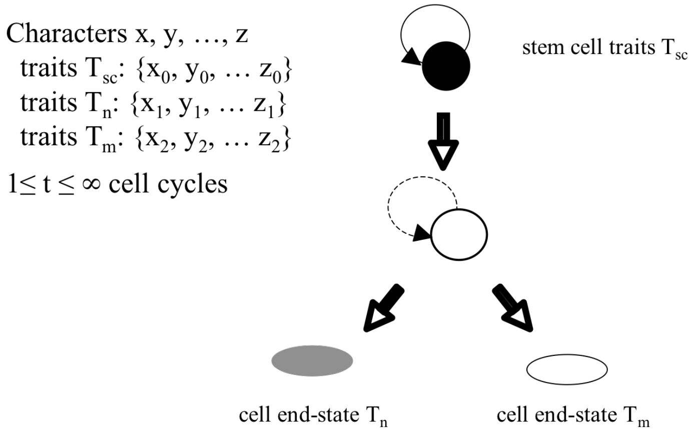
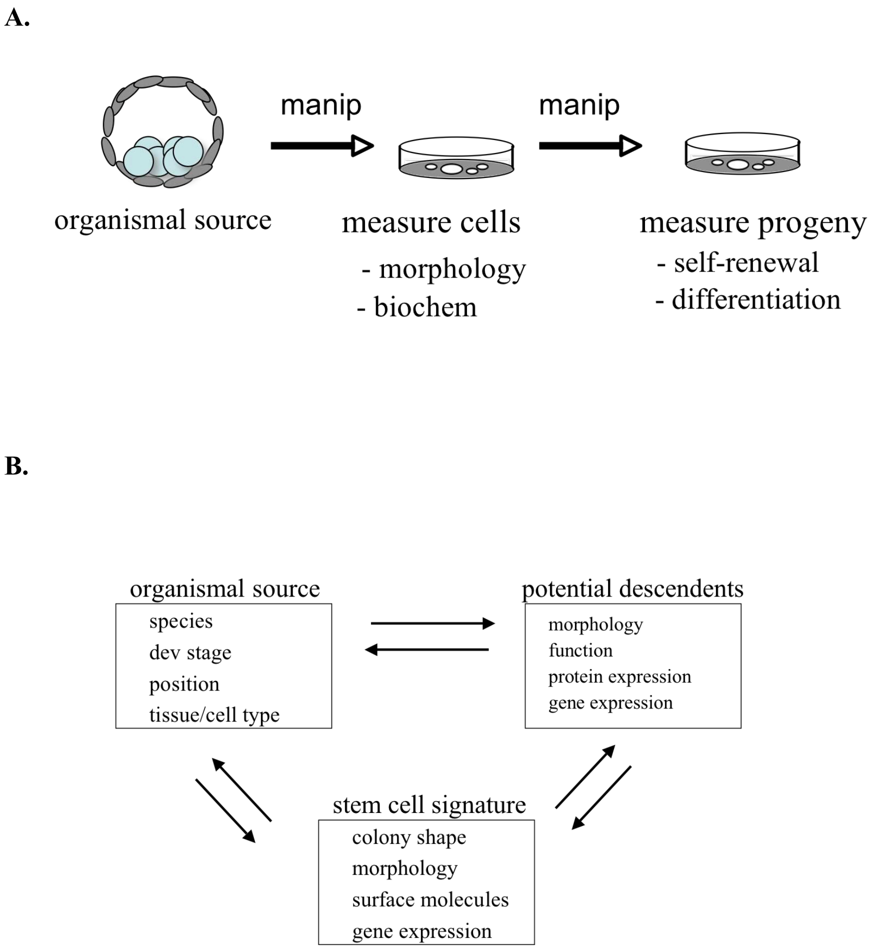

Characters x, y, …, z
traits T_{sc}: $\{x_0, y_0, \dots z_0\}$
traits T_n: $\{x_1, y_1, \dots z_1\}$
traits T_m: $\{x_2, y_2, \dots z_2\}$
$1 \leq t \leq \infty$ cell cycles
stem cell traits T_{sc}
cell end-state T_n
cell end-state T_m

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.



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.

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: $\mathbf{L}_{\mathbf{r}'}$ **= (r’-r)** × **p** = m (**r’**-**r**) × **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 $\boldsymbol{\mu}_{\mathbf{F/M'}}$ = (**r’**-**r**) × **F**. Then, deriving the angular momentum yields

$$\frac{d\mathbf{L}_{\mathbf{r}'}}{dt} = \frac{d((\mathbf{r}-\mathbf{r}')\times\mathbf{p})}{dt} = (\mathbf{r}-\mathbf{r}')\times\frac{d\mathbf{p}}{dt} + \frac{d(\mathbf{r}-\mathbf{r}')}{dt}\times\mathbf{p}$$

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**p/**dt = d(m**v) /** dt= m**a = F.** So finally, one gets

$$(1)\ \frac{d\mathbf{L}_{\mathbf{r}'}}{dt} = (\mathbf{r}-\mathbf{r}')\times\mathbf{F} = \mu_{\mathbf{F/M'}}$$

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

$$(1.5)\ \mu=\Sigma\ \mu_i = d\mathbf{L}/dt.$$

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**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 Σ 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 $(\mathbf{r'}\text{-}\mathbf{r}) \times \mathrm{d}\mathbf{p}/\mathrm{dt} = 0$, which is the case if $\mathrm{d}\mathbf{p}/\mathrm{dt}$ and $(\mathbf{r'}\text{-}\mathbf{r})$ 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 $\mathrm{d}\mathbf{p}/\mathrm{dt} = \alpha\ \mathbf{F}$, where α is real, not necessarily constant and not specified. Newton provides a more geometrical way to see the explanation:

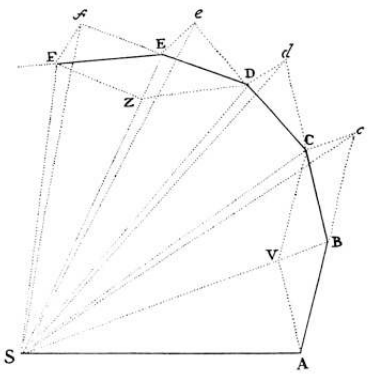


Figure 1: Geometrical demonstration of the Law of Areas by Newton (1726/1972)

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.

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

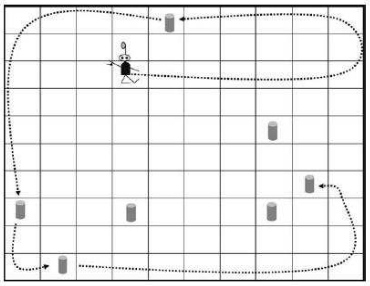
[3] Although we fully agree with Piccinini and Craver (2011) in that insofar as functional decompositions are explanatory, they are to be thought of as mechanism sketches and that the functional hypotheses are not independent of the question of mechanistic implementation.

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:
254355153256235251056355461151336154151034156110550150052030256256132252350325112
052333054055231255051336154150665264150266506012264453605631520256431054354632404
350334153250253251352352045150130156213436252353223135051260513356201524514343432
```

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 $(R \rightarrow S)$; 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 $R \rightarrow S$. 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 $(E \rightarrow H)$ 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 $(E \rightarrow H)$ 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 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).

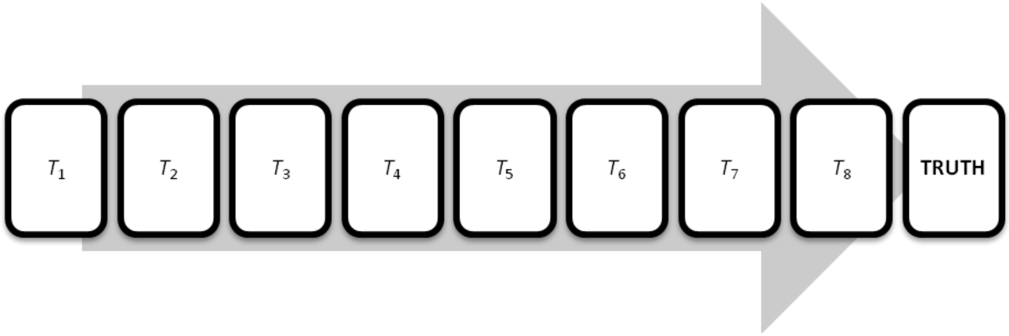
(R7) If the logical space of possible theories is vast, then there are no good reasons to believe that scientists have stumbled upon competing theories that are closest to the truth.

(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.

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

$$P_2/P_1 = \exp\left((S_2 - S_1)/k\right) \tag{1}$$

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 ln2 = 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 **x** and **π** 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

$$p(\lambda) = \text{constant.}\ Z(\lambda) \qquad (2)$$

where Z(λ) 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

$$F(\lambda) = -kT \ln Z(\lambda) \qquad (3)$$

treating Z(λ) as a partitition function, where the free energy is defined as

$$F(\lambda) = E(\lambda) - TS(\lambda)$$

Here E(λ) and S(λ) 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

$$p(\lambda_2)/\ p(\lambda_1) = \exp(-(F(\lambda_2) - F(\lambda_1))/kT) \qquad (4)$$

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]

$$X(\lambda) = - \partial/\partial\lambda|_T F(\lambda) = 0$$

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 $d = \partial/\partial\lambda|_T$, 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

$$p(\lambda_2)/p(\lambda_1) = \exp(-(-25kT)/kT) = \exp(25) = 7.2 \text{ x } 10^{10}.$$

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 $\int_V^{2V} kT/V'dV'$ = 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 ≥ 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.

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).

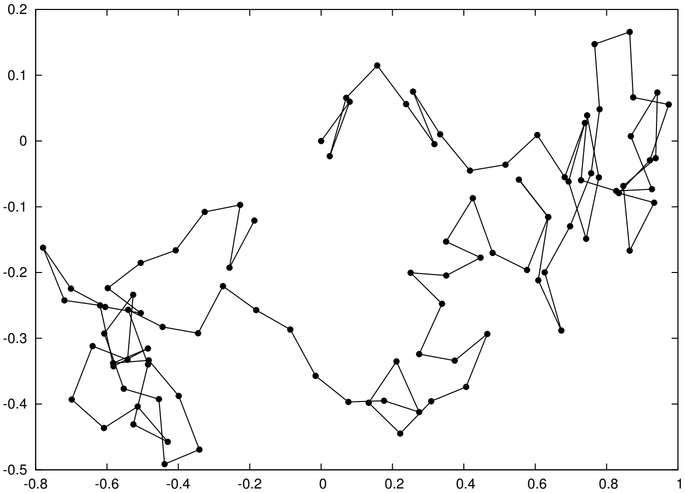


Figure 1: A simulation of a particle released at $(0, 0)$ undergoing Brownian movement. Inspired by Perrin's drawing of the Brownian motion of colloidal particles in water, viewed under the microscope (fig. 6 of Perrin, 1909, p. 81).

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):

$$m\frac{d^2x}{dt^2} = -6\pi\mu a\frac{dx}{dt} + X. \qquad (1)$$

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 ∅ 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.

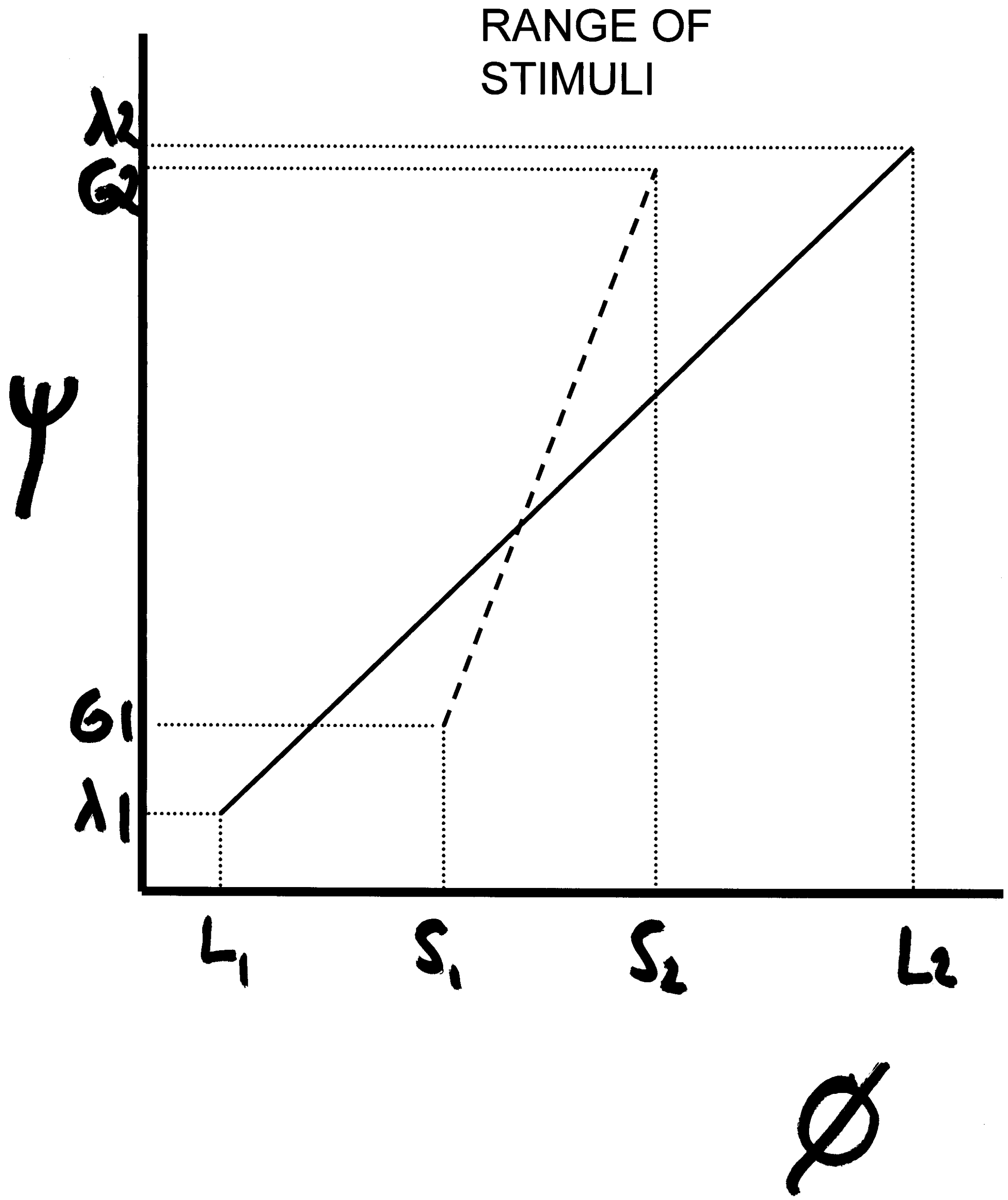
Papineau, D. (1994). The virtues of randomization. *British Journal for the Philosophy of Science*, 45(2), 437–450.

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

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.

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. .

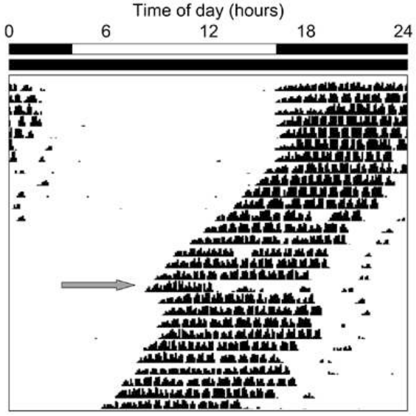


Figure 1. A basic actogram in which the top bar indicates a normal light-dark cycle for the first phase of the study (Days 1-7)_and constant darkness thereafter. The gray arrow identifies the day a light pulse was administered. (From http://www.photosensorybiology.org/id16.html.)

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.

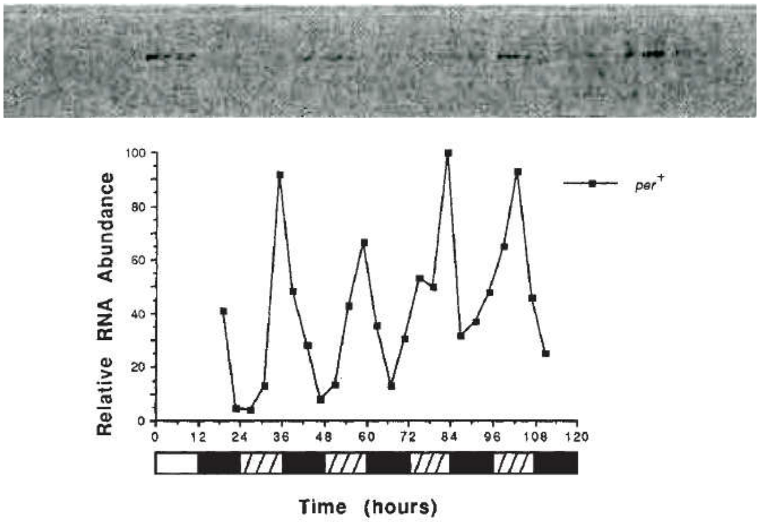


Figure 2. Two diagrams from Hardin et al.’s (1990) original portrayal of circadian oscillation in *per* mRNA levels in *Drosophila*. On top is a series of Northern blots (from different flies every 4 hours). Below this is a line graph of the same data. The Zeitgeber schedule is shown at the bottom, with 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.

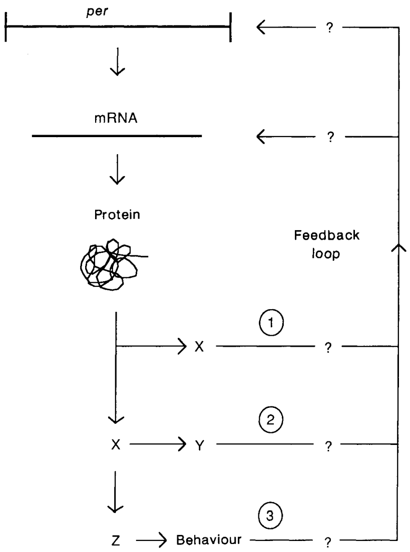


Figure 3. Hardin et al.'s (1990) representation of three versions of their proposed molecular mechanism for circadian oscillations in terms of a negative feedback loop. Question marks indicate points of uncertainty as to the origin and termination of the feedback operation.

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).

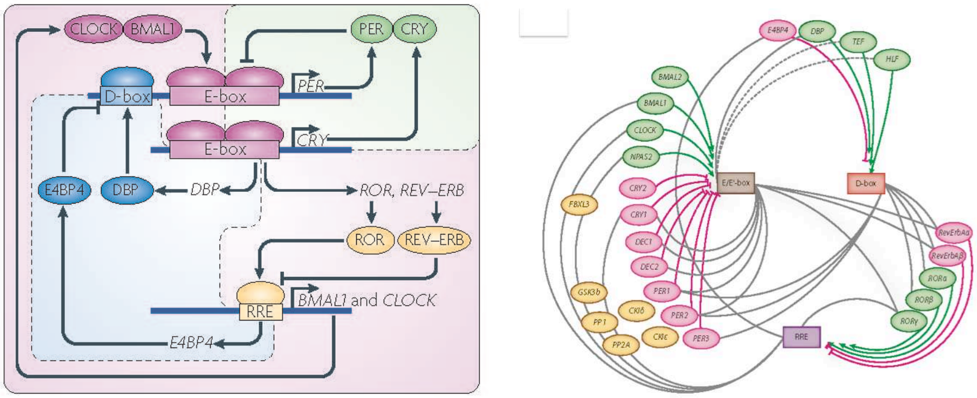


Figure 4. On the left is an example of a common way of representing the mechanism of the mammalian circadian clock, labeling genes in black italics and the proteins they express in colored ovals and using arrows to represent feedback loops (Zhang and Kay 2010). On the right an alternative representation (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.

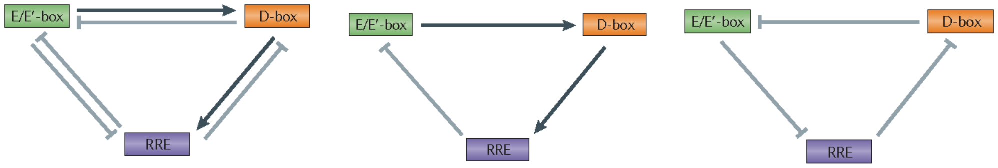


Figure 5. Hogenesch and Ueda's (2011) diagrams that abstract from the genes and proteins of the circadian oscillator to identify the basic causal circuit (left), which he then decomposes into two motifs (center and right) that are viewed as explaining the oscillatory behavior of the mechanism.

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.

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.

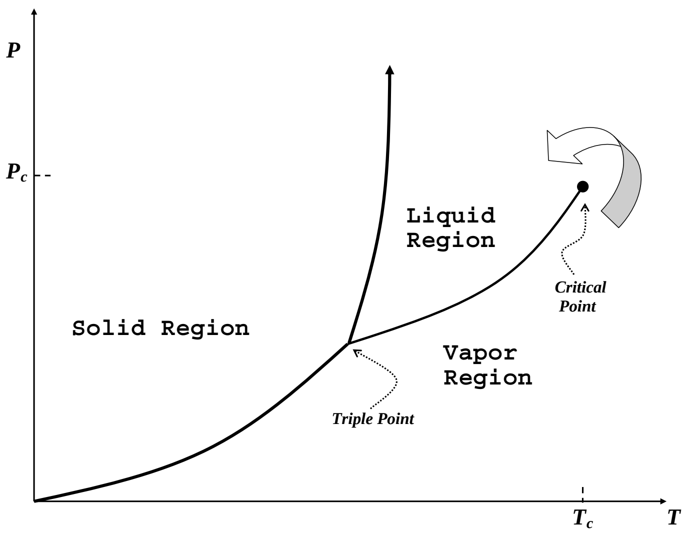


Figure 2.1: *Phase diagram of a generic material at fixed density.*

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

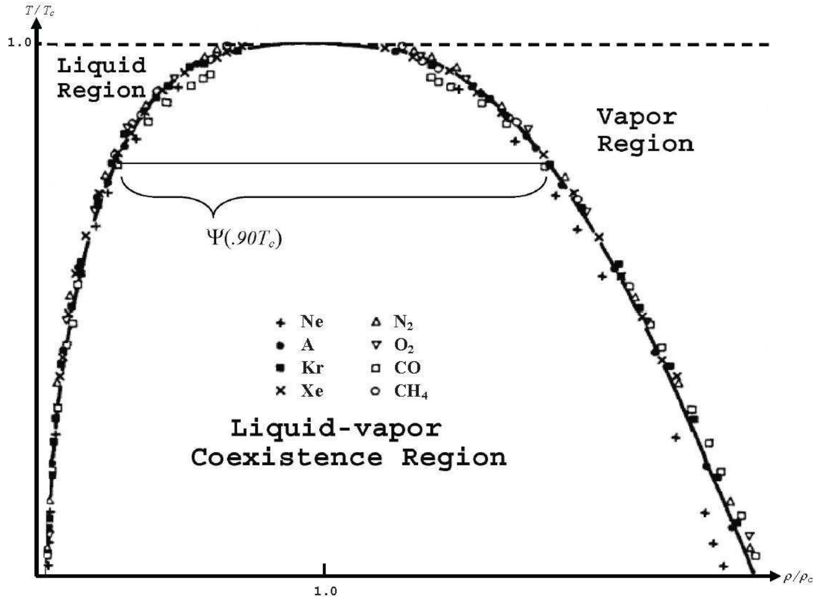


Figure 2.2: *Adapted plot of (Guggenheim, 1945) data rescaled for criticality.*

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 inevitably 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.

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.