**The London School of Economics and Political Science**

# Robustness, Evidence, and Uncertainty

## An exploration of policy applications of robustness analysis

Nicolas Wüthrich

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy, London, August 2017

# Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others.

I confirm that a shorter version of Chapter 2 appeared as *Wüthrich, Nicolas (2017): Conceptualizing Uncertainty – An Assessment of the Uncertainty Framework of the Intergovernmental Panel on Climate Change, In Massimi, M., Romeijn, J.-W., and Schurz, G. (eds.) EPSA 2015 Selected Papers. European Studies of Philosophy of Science: 105-117. Springer, Frankfurt.*

I confirm that a shorter version of Chapter 6 was co-authored with Wulf Gaertner and appeared as *Gaertner, Wulf and Nicolas Wüthrich (2016): Evaluating Competing Theories via a Common Language of Qualitative Verdicts. Synthese 193: 3293-3309.* I contributed to 70% of Chapter 6.



I declare that my thesis consists of 79085 words.

I can confirm that my thesis was copy edited for conventions of language, spelling and grammar by Andrew Mitchell, London School of Economics and Political Science, Language Centre.

Nicolas Wüthrich

# Abstract

Policy-makers face an uncertain world. One way of getting a handle on decision-making in such an environment is to rely on evidence. Despite the recent increase in post-fact figures in politics, evidence-based policy-making takes centre stage in policy-setting institutions. Often, however, policy-makers face large volumes of evidence from different sources. Robustness analysis can, prima facie, handle this evidential diversity. Roughly, a hypothesis is supported by robust evidence if the different evidential sources (such as observations or model results) are in agreement. In this thesis, I strengthen the case for the use of robustness analysis in evidence-based policy-making by answering open research questions about this inference technique. First, I argue that existing taxonomies miss a fruitful category of robustness reasoning, that is predictive stability. Second, I claim that derivational robustness analysis – the investigation of whether the results of different models are in agreement – can yield interesting insights even if not the entire relevant model space is covered by available models or if the model results are only partially in agreement. Third, I claim that expert knowledge is necessary to address questions that arise when one applies measurement robustness analysis – the investigation into whether multiple means of measurement yield the same result. Finally, I argue that, in situations where evidence from different measurements is not in agreement, it can be advisable to no longer take all of the evidence into account. This can be done in a rationally defensible way by choosing the most adequate theory or model underlying parts of the evidence set. I discuss examples from climate, medical, and economic policy-making to establish my claims.

*For my parents*
*Ursula and Hans Albert Wüthrich*

# Acknowledgements

First thank you to Roman Frigg, my supervisor, who guided me through this project with a great eye for the overall picture, with challenging but always encouraging feedback, and with a genuine interest in my work. I will certainly remember the countless exchanges in our peculiar language mix of Swiss German and English. I also want to thank Alex Voorhoeve, my second supervisor, who provided detailed feedback on my work in the later stages and who showed me how Philosophy teaching should look if it is supposed to be in touch with real world problems. The academic visit at the Federal Institute of Technology Zurich sponsored by Trude Hirsch-Hadorn provided the ideal setting for the final write-up of this thesis.

I am also thankful for the culture of the LSE Philosophy department. The exchanges with my fellow PhD students and faculty members during seminars and in the Ye Old White Horse pub were extremely stimulating. In particular, I want to thank James Nguyen, Jurgis Karpus, Mantas Radzvilas, Philippe van Basshuysen, Stuart Theobald, Aron Vallinder, Wulf Gaertner, Katie Steele, and Bryan Roberts for all their feedback. The regular Coca Cola breaks with Stephan Güttinger did not only teach me a lot about synthetic biology but helped me to navigate through the, at times, rough waters of the PhD life. Looking a bit further back, I would like to thank Dominique Kuenzle and Georg Brun for having equipped me with the tools needed to get started with analytical philosophy and all their encouragement.

Thanks to all the people at Advent Running who let me fall in love with ultrarunning and who, thereby, provided me with much needed balance. Finally, a big thank you to Martina Oehri, my fiancée, who supported me throughout this academic endeavour and who's presence truly made London my home.

# Contents

# Chapter 1

# Introduction: Types and Value of Robustness Analysis

Having to go to a hospital for treatment during the Middle Ages in Europe would have been a very different experience from today. Setting aside differences in levels of hygiene and infrastructure, the contrast in care one would have received is stark. A leading tool of diagnosis of that period was Galen's concept of the four humours, which was based on metaphysical beliefs from antiquity about the composition of matter. According to this view, the human body consists of four humours, namely blood, phlegm, yellow bile, and black bile. Health was marked as a state in which these four humours were in balance. An excess of blood, phlegm, yellow bile, or black bile made a person sanguine, phlegmatic, choleric, or melancholic (Hajar 2012, p. 159). Physical illnesses were said to be caused by imbalances in the four humours and, bloodletting in particular was seen as a common treatment to restore these imbalances (Hajar 2012, p. 159; Greenstone 2010, p. 12). Bloodletting remained a crucial treatment until the 19th century (Greenstone 2010, p. 12). Famously, George Washington was treated by having a significant amount of blood drawn after he had developed a fever and respiratory distress. He died the next night after this bloodletting (Greenstone 2010, p. 13). In hindsight, Washington's medical history does not come as a surprise. Amongst many others, studies by Pasteur, Koch, and Virchov in the 19th century showed

the harmful effect of bloodletting on patients (Greenstone 2010, p. 13) and eventually debunked the concept of the four humours.[1]

Elsewhere, in the 19th century, a large part of urban Britain was filled with what was called Victorian slums (Matthews 2010, p. 2). These slums were unplanned, densely populated areas with buildings in close proximity and no sanitary infrastructure. They were a product of population growth and rapid urbanisation due to the advancing industrialisation of Britain (Matthews 2010, p. 2). The growing awareness of the role of sanitary installations and the importance of hygiene towards the end of the 19th century led the UK government to intervene in the hitherto unregulated housing market (Johnson 2005, p. 20). The key interventions consisted in establishing minimal building standards (Boelhouwer and Hoekstra 2012, p. 363) and giving local councils the power to build homes (University of the West of England 2008, p. 3). However, the shortage of affordable houses was only slightly reduced by the council housing scheme since the councils built too few homes and the private sector did not jump in to fill this gap. Since the demand for housing was so high, private sector companies had no incentives to offer space at affordable prices (University of the West of England 2008, p. 3). The effectiveness of the council housing schemes of the late 19th century was curbed due to the market power of private sector construction firms. Hence, taking into account evidence about the market power and incentives of these private sector firms would have allowed the government to implement a more effective policy, for example, by building more governmental homes and by offering special renting schemes for the construction firms that would have allowed a reasonable payback period on their investments.

These two episodes from medicine and housing reveal that the use of evidence is crucial in designing effective interventions. But what kind of evidence should be gathered? And if evidence is gathered, how should this information be used to inform recommendations for interventions? These questions are particularly pressing if one looks at interventions at a larger scale such as policy interventions. These questions are the foci of an emerging field that has

[1]For an excellent in-depth treatment of the failures of medicine that followed Galen see Wootton (2007).

been named aptly *evidence-based policy-making* (see Cartwright and Hardie 2012; Montuschi 2009).

The *March for Science* movement, which calls on scientists and members of the public to speak out for a central role of science in policy-making, illustrates the idea behind evidence-based policy-making. Their mission statement reads as follows:

> The March for Science champions robustly funded and publicly communicated science as a pillar of human freedom and prosperity. We unite as a diverse, non-partisan group to call for science that upholds the common good and for political leaders and policy-makers to enact evidence-based policies in the public interest. (March for Science 2017, p. 1)

According to this mission statement, policy-makers who, for example, suggest a change to income taxation should pay attention to economic data on the effect of the taxation level on the government budget. The US Food and Drug Administration should consider the outcome of randomised control trials on the effectiveness of a new vaccine against measles. If school administrators want to change the cap on the number of students per class, they should consider educational research on the effect of teacher-student ratios on students' learning behaviour. Industrial policies that curb the rise in the global average temperature to two degrees Celsius compared to pre-industrial levels should originate from a careful understanding of the mechanics of the global climate.

The evidence-based medicine movement is another example that illustrates the idea of evidence-based policy-making. It is particularly illustrative since there is a high degree of self reflection amongst evidence-based policy-makers in medicine about their approach. Here is Sackett et al. (1996, p. 1)'s definition of the approach:

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research.

This view of policy-making as an evidence-based activity contrasts considerably with narratives regarding two recent political events: The United Kingdom's decision to leave the European Union and Donald Trump's election. These two events have been described as signs of the West having entered a *post-truth world* (The Economist 2016). This post-truth world is characterised by a declining effort to relate policy proposals to shared facts and to engage in open, reason-based exchanges. This thesis aims at strengthening evidence-based approaches to policy-making by addressing epistemological issues that surround evidential claims.

One such key issue is the fact that a large amount of potentially conflicting evidence from different sources can be relevant to a problem or hypothesis. Consider the case of class size regulation. Prima facie, evidence from psychology on the learning behaviour of children, economic evidence on cost implications on staffing levels, and sociological evidence on the correlation between access to private vs. public schools are relevant. Evidence often stems from multiple disciplines and supports conflicting hypotheses (Stegenga 2009). Douglas (2012, p. 139) coined the term 'complex evidence' for information of this sort. How should one go about policy-making in the face of such diverse evidence?

Looking at the discussion within the evidence-based policy domain, one tool emerges to address this issue of evidential diversity: *robustness analysis*. Robustness analysis is the investigation into how robust one's evidence is, that is, whether multiple ways of determining the truth of a hypothesis (e.g., different experimental techniques or model assumptions) are in agreement. Here are three examples of the view that robustness considerations are a valuable tool in assessing evidence for policy-making.

In an analysis of the relationship between trust and economic performance of a country, Beugelsdijk et al. (2004, p. 132) state the following:

> Our overall conclusion is that despite the variation in the size of the effect of trust on growth (...), our extensive robustness analysis further adds to the empirical evidence that trust matters for explaining variation in economic performance.

The Intergovernmental Panel on Climate Change (IPCC) – an organisation that synthesises state-of-the-art knowledge on the causes and effects of climate change – emphasises the centrality of robustness for the prediction and explanation of climate phenomena. To start, robust bodies of evidence are supposed to vindicate "high confidence" in explanations and predictions (Mastrandrea et al. 2010, p. 3). More specifically, agreement amongst different climate models, which is a form of robustness, is viewed as a contributing factor to increased confidence in climate predictions. The following is an example involving statements on changes in monsoon circulations (IPCC 2013, p. 1219, italics in the original):

> There is growing evidence of improved skill of climate models in reproducing climatological features of the global monsoon. Taken together with identified model agreement on future changes, the global monsoon, aggregated over all monsoon systems, is *likely* to strengthen in the 21st century (...).

In setting out the new research field of computational sociology – the attempt to understand social processes with the help of agent-based models – Macy and Willer (2002, p. 162-163) comment on the centrality of looking for robust results:

> Although simulation designs should use experimental rather than post-hoc statistical controls to identify underlying causal processes, that does not mean researchers should avoid statistical analysis. On the contrary, ABMs [Agent-based models], especially those that include stochastic algorithms, require replications that demonstrate the stability of the results.

In this introduction, I review the philosophical literature on robustness analysis. This literature has focussed on the use of robustness analysis in science, hence, the discussion might appear to the reader somewhat removed from the policy-making domain. However, the aim of this literature survey is threefold: First, this chapter introduces the conceptual resources on which I

will rely in this thesis. In particular, it will allow me to introduce the distinction between *measurement* and *derivational robustness analysis*. Second, it states clearly what I take to be the main arguments for the value of measurement and derivational robustness. This is a necessary step before I can turn to the application of robustness reasoning to questions in the policy-making domain. Third, I outline the specific research questions and the main claims of this thesis.

## 1.1 The starting point: Levins's and Wimsatt's discussion of robustness

In many contexts, scientists face an object of study that exhibits great complexity. Consider again the UK housing market. The dynamics in this market are dependent on the actions of current home owners, prospective buyers, the UK government and local councils, as well on a host of additional factors such as the health of the UK economy. These factors moreover influence each other in multiple ways. How should a scientist proceed when she faces the task of investigating such a complex system?

A natural suggestion would be to build a model that is maximally faithful to the complexity of the system. If one sought to predict the outcome of a first-time buyer support scheme, one would construct a model that captures the various interactions between actors in this market. Levins (1966, p. 421) calls this the brute force approach. Unfortunately, problems arise when one tries to build such maximally faithful models. To start, one faces a measurement problem. It is practically impossible to gather all the data to determine the parameters of the vast number of equations needed to specify the interactions between units in these complex systems (Levins 1966, p. 421, Weisberg 2006a, p. 629). Furthermore, these models will not have analytical solutions which is potentially detrimental for their use in explanations. An analytical solution is an explicit description of the dependencies between parts of the model. These dependencies allow for the identification of explanatory factors (Levins 1966, p. 421, Weisberg 2006a, p. 630).

Instead, Levins suggests using multiple, inaccurate models to explore a phenomenon.[2] However, this multi-model approach faces an immediate question: Since all of these models are inaccurate, how does one know which models make trustworthy predictions and which can be used in explanations? (Weisberg 2006b, p. 731)

Levins suggests examining the robustness of model results instead of selecting a single best model. Although his by now classic paper *The Strategy of Model Building in Population Biology* marks the starting point of a systematic treatment of robustness analysis, the notion of robustness had already been discussed earlier by philosophers of science.

Pierce (1868, p. 141) remarked the following about the philosophical method (my emphasis):

> Philosophy ought to imitate the successful sciences in its methods, so far as to proceed only from tangible premises which can be subjected to careful scrutiny, and to trust rather to the *multitude and variety of its arguments* than to the conclusiveness of any one.

When Whewell introduces the notion 'consilience of induction', he remarks that consilience

> takes place when one induction, obtained from one class of facts, coincides with an induction, obtained from a different class of facts. This consilience is a test of the truth of the theory in which it occurs. (as quoted in Laudan 1971, p. 369)

Implicit reference to a notion of robustness can also be found in Ayer (1956, p. 39), who discusses the belief formation process of a historian regarding past events (my emphasis):

[2]The two additional claims in Levins (1966) are that three modelling aims, generality, realism, and precision, cannot be maximised simultaneously and, hence, different modelling strategies are observable, depending on whether one aims at predicting or explaining a phenomenon. In the interest of the brevity of the introduction, I am setting these claims aside.

> (...) if these sources are *numerous and independent*, and if they agree with another, he [the historian] will be reasonably confident that their account of the matter is correct.

Now, according to Levins (1966, p. 423)

> (...) even the most flexible models have artificial assumptions. There is always room for doubt as to whether a result depends on the essentials of a model or on the details of the simplifying assumptions. (...) Therefore we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies.

Several points are worth noting. First, this account of robustness is formulated in terms of models. In fact, Levins (1966) discusses the issues of robustness exclusively in the context of population biology and its use of mathematical models. Second, Levins connects robust model results with truth. This led to a debate about how such a seemingly non-empirical form of confirmation can work. This debate will be taken up below. Third, Levins does not provide an argument for the link between robustness and truth. Instead, he provides examples of robust and non-robust theorem. He shows that the statement "in an uncertain environment, species will evolve broad niches and tend toward polymorphism" (Levins 1966, p. 423) is robust since it follows from three different models, that is, from a fitness set model, a calculus of variation argument, and a model specifying the genetic system. As an example of a non-robust theorem, he discusses the claim that "a high intrinsic rate of increase [the productivity of a population] leads to a smaller average population" (Levins 1966, p. 427).

Taking Levins (1966) as a starting point, Wimsatt (1981, p. 128) offers a characterisation of robustness analysis, or, what he treats as synonyms, methods of multiple determination or triangulation:

> (...) the variants and uses of robustness have a common theme in the distinguishing of the real from the illusory; the reliable from the unreliable; the objective from the subjective; the object of focus from artifacts of perspective; and, in general, that which is regarded as ontologically and epistemologically trustworthy and valuable from that which is unreliable, ungeneralizable, worthless, and fleeting.

According to Wimsatt (1981, p. 128), four steps characterise robustness analysis. First, a variety of independent derivation, identification, or measurement processes is conducted or analysed. Second, features that are invariant under these different processes are identified. Third, conditions are searched under which these invariances prevail. Fourth, failures of invariance are explained.

Wimsatt discusses a broader notion of robustness than Levins. In particular, additional modes of derivation (sensory modalities, measurement techniques, experimental procedures) enter the scene. Furthermore, he goes some way towards specifying the elements of robustness analysis by hinting at features that are shared by *all* types of robustness analysis.

In the next section, I review taxonomies of robustness analysis. The goal is to identify and characterise more precisely distinct types of inference rules falling under the broad umbrella of robustness analysis.

## 1.2 Types of robustness analysis

To the best of my knowledge, Woodward (2006) and Weisberg (2013) provide the most refined taxonomies of robustness analyses. As I argue below, the two taxonomies are complementary. According to Woodward (2006, p. 219), one must distinguish between four types of robustness analysis. I introduce these four types in this section. In Section 1.3, I discuss the conditions which need to be in place such that these techniques yield sound inferences. In these

two sections I report the state of the discussion in the literature.[3]

### 1.2.1 Inferential robustness analysis

Inferential robustness analysis comes into play when one faces an inference problem of the following kind (Woodward 2006, p. 219): There is a fixed body of data $D$ and a conclusion $S$ about the truth or falsity of a hypothesis. Inferences from $D$ to $S$ are only possible given background assumptions $A_i$. Background knowledge does not provide strong reasons to choose between different background assumptions. Inferential robustness analysis assesses whether $S$ follows under the different $A_i$. Accordingly, inferential robustness analysis gives a sense of the evidential warrant of a hypothesis.[4] Woodward (2006, p. 221-222) distinguishes between two ways of making this statement more precise. Using propositional logic, one can express the task of inferential robustness analysis as assessing the soundness of the following inference:

$$\frac{\begin{array}{c} A_1 \vee A_2 \vee ... \vee A_n \\ \forall i : (A_i \& D) \Rightarrow S \end{array}}{S}$$

Using a Bayesian framework, which allows ascribing probabilities to propositions, one can express the task of inferential robustness analysis as assessing the soundness of the following inference: Given the background assumptions $A_i$ contain all relevant assumptions and given that every background assumption from this set implies, together with the data $D$, that the hypothesis $S$ should be accepted (i.e., the degree of belief in $S$ is above a threshold level $c$), then the hypothesis should be accepted. This can again be put as an inference scheme:

[3]Note that this review does not cover the discussion of robustness in decision theory. To keep the scope of the thesis manageable, I am setting aside this literature on robust rules for decision making, that is, decision rules that deliver desirable outcomes even under severe informational constraints (see Hanson and Sargent (2001)'s robust control theory for an illustration).

[4]Orzack and Sober (1993, p. 541) discuss this form of robustness analysis under the heading "robustness within data sets".

$$\sum_{i=1}^{n} P(A_i|D) = 1$$
$$\frac{\forall i : P(S|A_i, D) \geq c}{P(S|D) \geq c}$$

Let me now turn to derivational robustness analysis.

### 1.2.2 Derivational robustness analysis

According to Woodward (2006), derivational robustness analysis comes into play when one investigates the stability of model derivations. A first sense of derivational robustness analysis can be formulated with respect to a *single* model $M$ that allows the derivation of some claim $P$. The model contains a set of assumptions $A$. Derivational robustness analysis investigates how sensitive the derivation of $P$ is given different assumptions $A_i$, that is, whether one can still derive $P$ given different $A_i$ (Woodward 2006, p. 231). Using propositional logic, one can express this sense of derivational robustness analysis as assessing the soundness of the following inference: Given the set $A$ contains all relevant assumptions and given that every assumption from this set implies, together with model $M$, the model result $P$, then $P$ is the case.[5] This can be put as an inference scheme:

$$\frac{A_1 \vee A_2 \vee ... \vee A_n \quad \forall i : (A_i \& M) \Rightarrow P}{P}$$

Woodward (2006, p. 231) also points out that one can investigate the stability of model derivations across *different* models. Instead of changing the assumptions of a single model, one investigates a set of distinct models. Models are distinct to the degree that their assumptions differ in a substantial way, for example, models describe phenomena on different levels or highlight

[5] I assume that one can speak of the same model even if one has changed some of the model's assumptions. As the following discussion will make clear, this liberal way of individuating models is in line with the discussion in the literature and the scientific practice.

different causal factors or mechanisms. In this case, one assesses the soundness of the following inference: Given the set of models contains all relevant models $M_i$ and every model of this set implies model result $P$, then $P$ is the case. This can be put as an inference scheme:

$$\frac{\begin{array}{c} M_1 \vee M_2 \vee ... \vee M_n \\ \forall i : M_i \Rightarrow P \end{array}}{P}$$

Weisberg (2006b, p. 737) provides a general framework which details the inference pattern of derivational robustness analysis. This framework allows separating different phases of derivational robustness analysis. According to Weisberg, four phases can be distinguished: i) Assessing whether a set of models allows deriving the same result (named the robust property); ii) if so, investigating whether the models share a common structure $C$ that allows deriving the robust property; iii) if so, combine steps i) and ii) to formulate a robust theorem with the general form "ceteribus paribus, if a [common structure $C$] obtains, then [robust property] will obtain"; iv) assessing under what conditions the relationship between $C$ and the robust property breaks down.

Weisberg (2013, p. 160-166) allows specifying step iv) further. He elaborates that three different ways exist of assessing the stability between the common structure $C$ and the robust property. To start, one can perform a *parameter robustness analysis*, determining if changes to a model's value of parameters changes the behaviour of a robust property.[6] Furthermore, *structural robustness analysis* involves changing the "mechanics of the model" (Weisberg 2013, p. 161). For mathematical models this involves adding new terms or altering interactions between existing terms in the model description. For material models this amounts to physically altering the model. Finally, *representational robustness analysis* holds the attributes of the models (i.e., its parameters and its structure) fixed but changes the way they are represented. This is the case when, for example, a material model is replaced by a

[6]This form of robustness analysis had already been noted by Orzack and Sober (1993, p. 540). Raerinne (2013, p. 289) classifies this type of analysis as sensitivity analysis.

mathematical model.

According to Woodward (2006, p. 233) derivational and inferential robustness analyses are distinct. In contrast to providing an assessment of the evidential warrant of a hypothesis in light of data, derivational robustness analysis starts not with data but with a well-defined model (or a set of models) and investigates the stability of deductions from this model (or set of models). In my view, it is unclear what this distinction between derivational and inferential robustness analysis amounts to in the case of the derivational robustness analysis of *a single model*. To start, it is unclear from a semantic point of view, since inferences and derivations subsume the same practices. Furthermore, looking at the formal representations of derivational and inferential robustness analysis provided by Woodward (2006) shows that the two patterns imply each other if one replaces the notion 'data' $D$ with 'model' $M$:

$$\frac{A_1 \vee A_2 \vee ... \vee A_n \qquad \forall i : (A_i \& D) \Rightarrow P}{P}$$

$$\frac{A_1 \vee A_2 \vee ... \vee A_n \qquad \forall i : (A_i \& M) \Rightarrow P}{P}$$

Hence, derivational robustness analysis can be viewed as encompassing inferential robustness analysis as a special case.

### 1.2.3 Measurement robustness analysis

Measurement robustness analysis comes into play when different measurement results are available to ascertain the truth of a hypothesis. The different measurement results are usually connected to questions about the reality (or artificial nature) of properties. The classic example is Jean Perrin's determination of the Avogadro number (Woodward 2006, p. 233; Cartwright

1991, p. 149). He determined the Avogadro number[7] via thirteen different methods, including the theoretical study of Brownian motion, the behaviour of radioactive bodies, and the movement of ions.

According to Woodward (2006, p. 234-235), there is a crucial difference between inferential and measurement robustness. In the case of measurement robustness, there are ideally independent measurement techniques available. These independent measurement techniques rely on assumptions which are not conflicting, that is, all of them can be true simultaneously. In the case of inferential robustness analysis, the different background assumptions usually conflict.

### 1.2.4 Causal robustness analysis

Finally, Woodward (2006) introduces the category of causal robustness analysis. Causal robustness analysis investigates whether a relationship between variables continues to hold if one manipulates a variable by intervention (Woodward 2006, p. 235). The idea behind assessing the stability of a relationship under investigation is an assumption about the nature of causality: causal relationships between variables, in contrast to mere correlations, remain stable under interventions on these variables (Woodward 2006, p. 235). There are multiple ways of making this idea more precise but most attempts share the view that the stability does not need to hold with respect to any conceivable intervention but solely with respect to those interventions that are relevant for manipulation and control (Woodward 2006, p. 235).

Woodward (2006, p. 235) gives an example from econometrics. Consider the following regression equation that connects the dependent variable $Y$ with a set of independent variables $X_i$ (with $a_i$ being the corresponding parameters) and an error term ($\epsilon_i$):[8]

[7]The Avogadro number is equal to $6.022140857 * 10^2$ atoms or molecules per mole of a substance.

[8]I have slightly simplified Woodward's example for the exposition of causal robustness analysis.

$$Y = \sum_{i=1}^{n} a_i X_i + \epsilon_i$$

Woodward (2006, p. 235) points out that if this regression equation captures a causal relationship between the dependent and independent variables, then this relation should be invariant under some range of manipulation on the independent variables $X_i$. The relation is invariant if the dependent variable $Y$ changes according to the regression equation given the manipulations of $X_i$.

## 1.3 Value of robustness

The introductory remarks have made clear that the tool of robustness analysis should be seen as valuable. However, there are prominent voices of criticism of robustness analysis. In their stage setting critique of Levins (1966), Orzack and Sober (1993, p. 539) put forward worries that derivational robustness analysis is a suspicious form of non-empirical confirmation:

> Should the fact that a given group is recognized within a variety of frameworks [species in a biological taxonomy] be grounds for increased confidence in its reality? It is worth considering the possibility that robustness simply reflects something common among the frameworks and not something about the world those frameworks seek to describe.

This exact worry is echoed by Sugden (2000, p. 22-23). He remarks the following:

> Notice how this mode of reasoning [robustness analysis] remains in the world of models (...). It makes inductive inferences from one or a small number of models to models in general (...). Obviously, however, it cannot be enough to stay in the world of models. If the theorist is to make claims about the real-world, there has to be some link between those two worlds.

Cartwright (1991, p. 154) voices a similar criticism with respect to econometricians applying derivational robustness analysis:

> 'Econometrician X used a linear form, Y a log linear, Z something else; and the results are the same anyway. Since the results are so robust, there must be some truth in them.' But (...) we know that at the very best one and only one of of these assumptions can be right. We may look at thirty functional forms, but if God's function is number thirty one, the first thirty do not teach us anything.

These remarks call for a careful discussion of the value of the different types of robustness. The aim of this section is to describe the current state of this discussion and to build the ground for identifying key open questions, which I am going to state in Section 1.4. The main claims of my thesis are responses to these open questions.

The literature has so far engaged with the value of derivational, measurement and inferential robustness. The main question of attention is whether robust model results or robust sets of evidence confirm results or hypotheses. Since there is to the best of my knowledge no literature on the value of causal robustness that has a similar focus and does not merge with the broader question of the nature of causality, I am setting this type of robustness aside here.[9]

### 1.3.1 The value of measurement robustness

Why is one justified in believing (or holding a high degree of belief in) a hypothesis if one is in a situation in which multiple evidential modes assert the truth of the hypothesis?

[9]For the sake of completeness, let me flag two further strands of discussion. To start, Woodward (2006, p. 231) points out that the failure of derivational robustness can generate new research questions. For example, if two models disagree about a prediction, an empirical test might be needed to differentiate between the two parameter values in the respective model description. Furthermore, Wimsatt (1987, p. 7-8) emphasises the value of robustness analysis as a research strategy. He claims that aiming for robust findings can, for example, provide a framework for a series of models of increasing realism and complexity and undercut the too ready acceptance of a preferred hypothesis by scientists.

The main argument for the confirmatory power of measurement robustness is akin to the no-miracles argument for scientific realism. Scientific realists claim that it would be a miracle to have theories that are empirically adequate to a high degree, if the entities, properties, and relations, that these theories postulated, did not exist (Putnam 1975, p. 73). In parallel, it would be a miracle, or, as Cartwright (1991, p. 149) puts it, a "coincidence" to obtain the same measurement result if there were not an underlying, causally active phenomenon generating the result (Culp 1995, p. 448; Bycroft 2009, p. 133).[10] In terms of its inferential logic, the no-miracles argument is an inference to the best explanation. The best explanation for the concurrence of measurement results is the existence of an entity that generates these measurement results (Eronen 2015, p. 3964; Bycroft 2009, p. 129).

Two attempts can be found in the literature to make this argument more precise. The first one delineates the domain of application of the no-miracles argument. If one is in the epistemic situation of knowing (or having an extremely high degree of confidence in) the reliability of a particular evidential mode (e.g., a satellite that allows for the measurement of cosmic background noise), then no further measurement technique is needed. As Cartwright (1991, p. 151) points out, measurement robustness is a guard against error in instruments and, hence, if one is confident in having a reliable instrument, then no guarding against errors is required. Accordingly, measurement robustness is not a necessary condition for establishing the reality of entities, properties, or relations (see also Eronen 2015, p. 3968; Schickore and Coko 2013, p. 302).

As a preliminary comment on the second attempt, let me add a note about the independence of measurements. To gain a handle on this concept, consider again the classical example of measurement robustness: the determination of the Avogadro number. As Cartwright (1991, p. 153) puts it:

[10]There exists also a slightly different version of this argument (see Woodward 2006, p. 234; Cartwright 1991, p. 150): If the measurement techniques are independent, then it is reasonable to assume that all techniques are vulnerable to different kinds of errors. If so, then a huge coincidence would be necessary for multiple errors to be operating at the same time to have produced a wrong result. Hence, if all techniques agree, it is reasonable to assume that they agree due to measuring a real phenomenon and not an artifact.

> The various different instruments involved [in determining the Avogadro number] use assumptions which are, in the best case, wholly independent of each other. (...) [The] independent instruments [are] doing different things (...).

Cartwright (1991, p. 149) further elaborates that the strength of the measurement support for the Avogadro number comes from the fact that different experimental procedures, with associated different experimental situations, skills, and assumptions, lead to the same result. This role of independence is also mentioned by Woodward (2006) and Culp (1995). Woodward (2006, p. 234) further illuminates the kind of independence involved here by stressing differences in instrument design, operation according to different causal principles, and different assumptions necessary to interpret data. Culp (1995, p. 454) identifies three factors for determining the degree of measurement robustness of a body of data: i) the size of the set of techniques producing comparable data; ii) the degree to which different measurement techniques rely on different assumptions to interpret the raw data; iii) for each measurement technique, the degree to which its assumptions are dependent on the theory of the object of measurement.

Let me now turn to the second attempt to make the no-miracles argument more precise. It is Bayesian in nature. Why is diverse evidence more valuable than evidence from a single source? The key task this attempt faces is spelling out a notion of independence in probabilistic terms that is not prone to counter-examples. Full-blown statistical independence is too much, since there can be correlation between measurement devices that are different in important ways (e.g., thermometers that are based on different physical principles, such as mercury or air pressure, give readings that are statistically correlated) (Eronen 2015, p. 3969). Franklin and Howson (1984, p. 52) define independence in terms of less than perfect correlation between measurement techniques. Two measurement procedures are less than perfectly correlated if the respective results produced from the same input data are less than perfectly correlated. Roughly, results are less than perfectly correlated if there is a difference in the posterior probability value of two measurement results

from different procedures given a series of the same previous measurement results. For two test procedures $E$ and $E'$ and corresponding series of results $e_1, e_2, ..., e_n$ and $e'_1, e'_2, ..., e'_m$, less than perfect correlation of results can be defined more precisely as follows. If and only if, for all $m > m_0$, for some $m_0$:

$$P(e_{m+1}|e_1 \wedge e_2 \wedge ... \wedge e_m) > P(e'_{m+1}|e_1 \wedge e_2 \wedge ... \wedge e_m)$$

and for all $n > n_0$, for some $n_0$:

$$P(e'_{n+1}|e'_1 \wedge e'_2 \wedge ... \wedge e'_n) > P(e_{n+1}|e'_1 \wedge e'_2 \wedge ... \wedge e'_n)$$

then the two experimental procedures $E$ and $E'$ are less than perfectly correlated (Franklin and Howson 1984, p. 52).

If these two conditions hold, then it can be shown that the marginal increase of the posterior probability of a hypothesis is larger if the additional experimental result comes from a different experimental procedure compared to the case in which the additional experimental result comes from the same experimental procedure (Franklin and Howson 1984, p. 52). Without going into the details of alternative attempts to flesh out this Bayesian analysis[11], let me note a problem with this formal account of independence. As Collins (1984, p. 172) points out, in many episodes of experimental scientific practice, scientists have disagreed substantially about the (dis)similarity of experimental procedures.[12]

Schupbach (2016) has recently proposed an alternative Bayesian account of evidential diversity. Schupbach (2016, p. 13) defines evidential diversity in terms of alternative explanations for the evidence that are ruled out. A set of evidence is more diverse than another, if it was generated by means of determination that allowed the ruling out of more alternatives or competing

[11]See Fitelson (2001) and Sober (1989) for further discussion.

[12]The discussion of formal accounts of independence is kept deliberately short here. For example, see Schupbach (2016, p. 6-12) for a discussion on reliability and confirmational independence. Reliability independence defines independence as the situation in which each means of determination is or is not reliable independent of each other. This notion underlines Wimmsatt's chain argument for measurement robustness: A linear chain of justification cannot be stronger than its weakest link. In contrast, a web of independent lines of justification is no weaker than its strongest member (Schupbach 2016, p. 8).

explanations. He defends this criterion, which he makes formally precise, by showing that it captures the ongoing scientific practice and that it is normatively compelling, that is, it accounts for the special normative status that is ascribed to robustness analysis (Schupbach 2016, p. 6).

Stegenga (2009) is not convinced by these attempts to strengthen the argument for the value of measurement robustness analysis. To start, he argues that it is unclear how to individuate different modes of evidence (Stegenga 2012, p. 14-16). In particular, existing accounts of separating evidential modes are not satisfactory: Culp (1995)'s stressing of theory-ladenness is problematic, since not all evidence is theory-laden to the same extent and one can have one theory but still two different modes of evidence. Reference to all background assumptions does not do the trick either since it is not clear what background assumptions are relevant for establishing independence. Reference to the problematic background assumptions requires that one has a way of identifying what the problematic background assumptions are in the respective measurement techniques. In addition, concordant multi-modal evidence does not necessarily give the correct answer. To put it bluntly, coincidences can occur and wrong inferences about the existence of an entity can be the consequence. Stegenga (2009, p. 652-653) mentions the case study of the mesosome in cellular biology, where electron and light microscopes supported the existence of these cell components.[13] Finally, and most importantly in Stegenga's view, multi-modal evidence is usually not concordant, that is, robust, but the different modes of determination point in different directions. It is far from clear what one should do in these situations and, according to him, robustness analysis is worthless in providing much-needed guidance (Stegenga 2009, p. 654). He distinguishes between two types of discordant evidence: discordant evidence can either be inconsistent, when the evidence bears on the very same hypothesis (e.g., Petri dish suggests $x$ and test tube suggests $\neg x$); or incongruent, when the background assumptions in the different evidential modalities do not cohere and, consequently, the modes produce evidence statements formulated in two non-translatable languages (Stegenga 2009, p. 654).

[13]Interpretation of this case study is hugely contested. See Culp (1994) for a discussion.

Hey (2015, p. 70-71)'s response to Stegenga helpfully clarifies the issue at stake here. First, if evidence is discordant, this might simply be because different modalities are capturing different mechanisms at play (Hey 2015, 60). Second, one must distinguish between the method of robustness analysis and the state of having robust results (Hey 2015, 70-71). Although robustness analysis cannot tell us what to believe in situations of discordant multi-modal evidence, it helps to generate new hypotheses that one can investigate. So, to be precise, the debate concerns what to believe in the situation of discordant multi-modal evidence regarding the working of the same mechanism underlying a phenomenon.

### 1.3.2 The value of derivational robustness

Why is one justified in believing (or holding a high degree of belief in) a hypothesis when a class of models implies this hypothesis? A discussion on the confirmatory import of derivational robustness analysis has occupied the largest part of the literature on robustness analysis.

The starting point of the discussion was marked by Orzack and Sober (1993). They argued that derivational robustness analysis is a type of *non-empirical confirmation* and, hence, that it is suspicious (Orzack and Sober 1993, p. 544). They start from the formal representation of derivational robustness analysis which involves multiple models (recall that $M_i$ denotes the $i$th member of a set of models and $P$ a model derivation):

$$\frac{\begin{array}{c} M_1 \vee M_2 \vee ... \vee M_n \\ \forall i : M_i \Rightarrow P \end{array}}{P}$$

According to Orzack and Sober (1993), exploring the first premise ($M_1 \vee M_2 \vee ... \vee M_n$) shows that derivational robustness analysis does not have confirmatory power. The inference from the two premises to the conclusion is sound if and only if one knows that at least one of the models $M_i$ is true. In practice, this is not the prevalent epistemic state (see also Parker 2011, p. 583). If one knows that all of them are false, then robustness of the result

across the models does not establish and needs not support the truth of the conclusion. If one is unsure whether one of the models is true, then one again has no assurance that the conclusion is true (Orzack and Sober 1993, p. 538-539).

The problem with this objection is that this formal representation of derivational robustness analysis is a shorthand in an important respect. Each model $M_i$ must be understood as a set of assumptions. This set of assumptions can contain true and false statements. Hence, although a model might be a literally false description of a target system (because it contains at least one false assumption), there can still be a substantial overlap of the respective true assumptions of the models. Accordingly, it is not prima facie clear that one cannot gain some insights from the robustness of a derivation from a set of literally false models. In particular, this would be the case if the falsehoods in the models could be understood as irrelevant falsehoods with respect to a target system.[14] The upshot here is that, in the context of derivational robustness analysis, one should not be ascribing truth values to entire models but to model assumptions (Weisberg 2006b, p. 733; Levins 1993, p. 553).[15]

To make the subsequent discussion more concrete, let me introduce an example of a model result that is widely accepted as being robust: the Volterra property in population biology.

The Volterra property was derived with the help of the Lotka-Volterra model of predation. Predation is an extensively studied phenomenon in ecology since it is one of the key forces that keeps a population under the carrying capacity of an environment (Weisberg and Reisman 2008, p. 108). Put simply[16], the model consists of two negative coupled differential equations describing the evolution of the prey and predator populations respectively. The model depicts this evolution as an undampened oscillation. Accordingly, the prey and predator population grow out of phase with each other, that is, an increase in the prey population increases the number of predators which

[14] I am going to elaborate on this point below.

[15] Interestingly, in a recent contribution Hands (2016, p. 44) suggests that derivational robustness analysis can increase our confidence in models as a whole. However, he provides no argument for why the confirmatory import should be attributed to models as a whole.

[16] The model is introduced in more detail in Chapter 4.

in turn drives down the number of prey (Weisberg and Reisman 2008, p. 112). Given this model, one can derive algebraically the Volterra property which states that "a general biocide, any substance that has a harmful effect on both predators and prey, will increase the relative abundance of the prey population." (Weisberg and Reisman 2008, p. 113). The Volterra property can be derived from a wide range of parameter values for the two differential equations (Weisberg and Reisman 2008, p. 115), from structurally different models (e.g., a model with an upper boundary for the environment's carrying capacity, Weisberg and Reisman 2008, p. 119), and models that use a different way of representing the attributes of the model (e.g., an agent-based computational model of the coupled predator-prey interaction, Weisberg and Reisman 2008, p. 128). Then, the key question is whether the fact that the Volterra property is a robust model result confers confirmation onto the Volterra property? Three preliminary remarks can make this question more precise.

First, there are clear cases in which this question is irrelevant because there are other sources of confirmation that are providing the confirmation for a hypothesis by themselves. If one has such confirming evidence for the property, for example in the form of observations about biocides in actual predator-prey ecosystems, then derivational robustness analysis is beside the point with respect to the question of confirmation of the property (Lehtinen 2016, p. 1; Orzack and Sober 1993, p. 541). Hence, the question should be re-formulated as follows: In the case of no direct evidence for the hypothesis of interest, does derivational robustness confer confirmation onto this hypothesis?

Second, one confirmatory function of derivational robustness analysis is widely accepted in the literature. It involves confirmation in a rather trivial sense. What derivational robustness analysis does confirm are claims about the relative importance of various assumptions of a model with respect to the model result of interest (Kuorikoski et al. 2010, p. 543; Muldoon 2007, p. 882; Odenbaugh and Alexandrova 2011, p. 765; Weisberg 2006a, p. 643; Weisberg and Reisman 2008, p. 106). Take the Lotka-Volterra model for example. It turns out to be the case that if one introduces a carrying capacity of the envi-

ronment, the property of the undamped oscillation disappears (Weisberg and Reisman 2008, p. 118). Hence, derivational robustness analysis allows determining that assumptions about the effect of population density are relevant for the property of undamped oscillations.

Third, Lisciandra (2016, p. 7) highlights the distinction between de-idealisation and derivational robustness analysis. In a de-idealisation model assumptions are replaced with more realistic assumptions. It is not necessarily the case that in a process of derivational robustness analysis assumptions are replaced by more realistic ones (see also Odenbaugh and Alexandrova 2011, p. 759; Kuorikoski et al. 2010, p. 558). In the case of the Volterra property, it might be that an agent-based representation of predator prey interactions contains the same number of, or even more, idealisations (or idealises certain aspects of this interaction to a higher degree). Hence, one cannot presume that derivational robustness analysis confers a degree of confirmation onto robust properties via a process of approximating a more realistic model of a phenomenon of interest (Lisciandra 2016, p. 8). Consequently, the question about the confirmatory import of derivational robustness analysis in its sharpest form can be put as follows: In the case of no direct evidence for the hypothesis of interest and the absence of a de-idealisation of the model under investigation, does derivational robustness confer confirmation onto the robust hypothesis?

Broadly speaking, there are two arguments in the literature that support an affirmative answer to this question. I will discuss them in turn.

### The first argument: Confirmation via direct comparison

Kuorikoski et al. (2010) argue that derivational robustness analysis in itself cannot confer confirmation onto robust model results. A comparison between the model and empirical information about a target system is needed (Kuorikoski et al. 2010, p. 549, 551; see Forber 2010, p. 38 for a defence of the same claim). However, they argue that derivational robustness analysis can play a vital role in such an empirical process of confirmation. To see this, they differentiate between three types of model assumptions: Substantial

assumptions, which describe a set of causal factors or mechanisms; Galilean assumptions, which isolate the working of the causal factors or mechanisms by idealising away additional causal factors; and tractability assumptions, which are assumptions that are empirically not well motivated but which facilitate the derivation of results (Kuorikoski et al. 2010, p. 547).[17] According to Kuorikoski et al. (2010, p. 548), in derivational robustness analysis you aim to show that the result of interest (e.g., the Volterra property) is driven by substantial and not by tractability assumptions. If there is evidence for the truth of the substantial assumptions with respect to a target system, then the robust model result is confirmed (Kuorikoski et al. 2010, p. 552; see also Kuorikoski and Lehtinen 2009, p. 127).

Lehtinen (2016, p. 5) defends a more refined version of this claim. He argues that this type of confirmation does not only occur if one has direct evidence for the substantial assumptions, but also in the case of indirect evidence, that is, evidence supporting the assumptions that allow deriving both this evidence claim and the robust model result. According to this first approach to the confirmatory power of derivational robustness analysis, "there is nothing dubiously non-empirical or Münchhausen-like in the epistemic import of robustness analysis" (Kuorikoski et al. 2010, p. 551). Justus (2012, p. 800-801) helpfully captures this first approach as follows: "robust theorems establish conduits through which empirical support for $C$ [the substantial assumptions driving the result] can transmit to $R$ [the robust model result]."[18]

Odenbaugh and Alexandrova (2011, p. 763-764) take issue with this approach. They claim that showing that one has arrived at a theorem that is derivationally robust with respect to the tractability assumption is not suffi-

[17]This tripartite is not as clear-cut as it might seem. The same assumption can fall into different categories given different epistemic contexts. For example, the assumption of zero transportation costs can be a Galilean idealisation in a model of international trade. The very same assumption can also be a substantial assumption if one models the informational transmission capacities of computer networks (see Musgrave 1981's influential discussion of this point in relation to Friedman's instrumentalism and Hands 2016, p. 38).

[18]Weisberg draws attention to an underlying assumption of this view of the relationship between robust theorems and target systems. One needs to assume that one can represent the dynamic of a system via mathematical functions (Weisberg 2006b, p. 740-741; Weisberg 2013, p. 168). He calls this "low level confirmation" without which one could not even connect model results to target phenomena.

cient to lead to an adequate representation of the causal mechanism operating in a target system. For if you replace one tractability assumption with another idealised one, it might be the case that no causal relationship is represented. To represent a causal relationship, all the false assumptions that go into the formulation of the robust theorem must be replaced with true ones (Odenbaugh and Alexandrova 2011, p. 764). Consider the Volterra property again, but this time embedded in the Volterra principle (Weisberg 2013, p. 159):

> Ceteribus paribus, if a two-species, predator-prey system is negatively coupled, then a general biocide will increase the abundance of the prey and decrease the abundance of predators.

What one needs to show according to Odenbaugh and Alexandrova (2011, p. 764) is that there is some set of true auxiliary assumptions (involving tractability and Galilean assumptions) under which the relationship between coupled predator-prey systems and a general biocide holds. According to them, unless one specifies correctly a causal mechanism, the supposed increase of confidence should not take place. In particular, they argue that it is not enough to show, as Kuorikoski et al. (2010, p. 561-562) claim, that the assumptions that allow one to derive this relationship are independent in the sense that they are unlikely to give rise to the same failure of inference. Odenbaugh and Alexandrova (2011, p. 762-763) are not convinced, since this explicit reference to one version of the no-miracles argument in defence of measurement robustness overlooks that models are substantially different from experimental modes of determination (see also Eronen 2015, p. 3965).[19]

Odenbaugh and Alexandrova (2011)'s conclusion is that derivational robustness should not be regarded as a tool of confirmation. Rather, they see derivational robustness analysis as a method of discovery. For derivational robustness analysis allows refining what they call "open formulas" by updating conditions under which the purported causal relationship holds. Open

[19]Lisciandra (2016, p. 11-14) raises an additional worry. It might be the case that one cannot change tractability assumptions in isolation. Since tractability assumptions are introduced to make the mathematics work, they are likely to come as interrelated sets of assumptions.

formulas are rough templates for causal claims and can be expressed as follows: "In a situation $x$ with some characteristics that may or may not include [conditions] $(C_1, ..., C_n)$, a certain feature $F$ causes a certain behaviour $B$" (Odenbaugh and Alexandrova 2011, p. 769-770). Accordingly, derivational robustness analysis disciplines our background knowledge in a formally precise way and facilitates the empirical confirmation of causal claims (Odenbaugh and Alexandrova 2011, p. 770).

Let me now turn to the second argument that ascribes confirmatory power to derivational robustness. In my view, it is this argument that is the target of the prominent attacks of Orzack and Sober (1993) and Sugden (2000) on derivational robustness analysis.

### The second argument: Confirmation via covering the space of possibilities

Despite stressing the fact that there cannot be confirmation through derivational robustness analysis without direct empirical confirmation of parts of a robust theorem, one can find passages in Weisberg that suggest confirmation without such direct empirical information. Here is the passage that I have in mind (Weisberg 2006b, p. 739):

> Explaining a real-world phenomenon or predicting its occurrence requires us to know that the common structure is actually being instantiated and that no other causal factor is preempting the efficacy of the common structure. One way to determine if the common structure is being instantiated and if any preempting causes are present is to conduct an empirical investigation. While this is the most reliable way to ensure that a robust theorem can be applied, it is often impractical or impossible to collect the relevant data. In fact, robustness analysis is usually introduced in situations in which data are hard to obtain. Fortunately, there is an alternative that, while not completely reliable, can give us good reasons to believe the predictions and explanations of robust theorems.

Weisberg claims that this alternative involves answering two questions that robustness analysis can partially answer: "1. How frequently is the common structure instantiated in the relevant kind of system? 2. How equal do things have to be in order for the core structure to give rise to the property?" (Weisberg 2006b, p. 739). According to Weisberg, the first question can be answered if one assesses how diverse the set of models is that are considered and that share the common structure. If this set is sufficiently diverse, then "it is very likely that the real-world phenomenon has a corresponding causal structure" (Weisberg 2006b, p. 739). Unfortunately, Weisberg does not give an account of the heterogeneity or diversity of models. As Orzack and Sober (1993, p. 539) make clear, this question is not trivial. Being able to say whether a set of models is sufficiently diverse presupposes a handle on the question of how the dependency relation between models should be expressed. In Orzack and Sober (1993)'s view, two salient attempts to make this notion more precise fail. First, model independence cannot be understood as logical independence. Logical independence of two propositions is given if neither of the two implies the truth or falsity of the other. The problem is that in applications of derivational robustness analysis one looks at competing models (involving contradictory assumptions) (see also Cartwright 1991, p. 154; Bycroft 2009, p. 136; Schupbach 2016, p. 7). Second, it is problematic to understand model independence in terms of statistical independence since for this one needs to be able to specify the space of models, ensure that the models can be viewed as a non-overlapping partition of this space, and ascribe probabilities to the elements of this space. If one is not a Bayesian, the last task might already pose a problem (Orzack and Sober 1993, p. 539).

The second question, Weisberg claims, can be answered by conducting a stability analysis of the robust theorem. Since this is already part of Weisberg's four step process of robustness analysis, this information is readily available. If the robust theorem is stable, that is, the link between the common structure and the result persists against perturbations, then one can expect the effectiveness of the causal structure in the target system (Weisberg 2006b, p. 740).

Although this idea is no longer present in his most recent contribution to

the topic (see Weisberg 2013), the alternative view that derivational robustness analysis can be confirmatory without direct evidence for the common causal structure in a robust theorem has been picked up by others (see Justus 2012, p. 801). In particular, Parker (2011)'s discussion of the significance of robust model predictions can be viewed as a contribution to this question.

Parker (2011) asks under what conditions robustness of model results have special epistemic merit in the sense that there is a higher likelihood of truth, increased confidence, or increased security of the hypothesis upon which all the models agree. I will not discuss her leading case study – multi-ensemble climate modelling – since this will be taken up in Chapter 2. The upshot of her paper is that model agreement does not justify a higher likelihood of truth, increased confidence, or increased security in the hypothesis. Model agreement on a hypothesis $H$ in the case of climate models does not increase the likelihood of truth of $H$ (Parker 2011, p. 584), because the ensemble of models is an "ensemble of opportunity" and does not accurately reflect structural model and parameter uncertainty (Parker 2011, p. 585). Model agreement on $H$ also does not justifiably increase our confidence in $H$ given $E$. Given a Bayesian analysis, confidence in $H$ increases if and only if $P(E \mid H) > P(E \mid \neg H)$. Since one only has a limited understanding of the climate systems, and hence, of the limitations of the state of the art climate models, one cannot give an argument for why this inequality must hold (Parker 2011, p. 591). Furthermore, a sample-based argument for this inequality also fails (Parker 2011, p. 593). Assume you have a set of climate models that each satisfy a quality threshold criterion. Now draw a random sample from this set and evaluate what fraction of these models indicates the truth of $H$. This fraction then is a proxy for the fraction of models in the initial set that indicate the truth of $H$. The problem with this approach is again that the ensemble is a not a random sample but an ensemble of opportunity (Parker 2011, p. 594).

These problems point towards an assumption that needs to be in place for non-empirical derivational robustness having confirmatory import: the models that are evaluated in the analysis need to span the space of possible models. Possible models is a vague notion here but it can be made – in a

first step – more precise as follows: The space of possible models can be spanned by exploring possible model structures (i.e., alternative substantial assumptions) and possible parameter values. Without such an assumption in place, it is not clear that one has answered convincingly either one of the two questions raised by Weisberg (2006b). Making this assumption explicit, immediately prompts the question as to what happens to the epistemic import of derivational robustness analysis if only a subset of this model space can be explored (see Weisberg 2006a, p. 641 for a similar point). In his response to Orzack and Sober, Levins (1993, p. 553) suggests that the more of the possibility space is covered, the higher the confidence should be in the robust model result. However, Levins mentions this only in passing and does not give a detailed argument for this claim.

### 1.3.3 The value of inferential robustness

Why is one justified in believing (or holding a high degree of belief in) a hypothesis when one is in a situation in which multiple ways of inferring a hypothesis from a data set are in agreement?

Recall that Woodward (2006) provided two formulations of the inference pattern (see Section 1.2.1). These formulations suggest that two conditions need to be in place such that the link between robustness and truth or high probability of a hypothesis is warranted (Woodward 2006, p. 221):

**Completeness condition** The set of auxiliary assumptions $A$ contains the true auxiliary assumption $A_k$.

**Robustness condition** The evidence together with every individual element of the set of auxiliary assumptions allows inferring the hypothesis $S$ or confers a posterior probability onto $S$ that is above a specified degree of belief threshold.

Taken together, these two conditions provide sufficient conditions for the link between robustness of a result and its truth (or high probability) (Woodward 2006, p. 221). However, it is unclear what happens to the value of

the inference patterns if one relaxes the completeness or robustness condition (Woodward 2006, p. 222-223). As Woodward (2006, p. 222) points out, in many real-world applications of robustness analysis, one or both of these conditions fail to hold.

Let us assume that completeness holds but robustness fails. In this scenario, only a few auxiliary assumptions allow inferring a particular hypothesis (or confer the required degree of belief in its truth). Woodward (2006) seems to suggest that our confidence could increase in parallel to the size of the set of auxiliary assumptions that allow inferring the result (or confer the required degree of belief in its truth). However, he quickly adds that this inference is only warranted if the probability mass is not concentrated on a few auxiliary assumptions for which the robustness condition does not hold.

Or consider the case in which robustness holds but completeness fails. This case is not discussed by Woodward (2006). However, one can shed light on this case by extending the parallel to the scenario of derivational robustness analysis. Accordingly, the criticism of Orzack and Sober (1993) is also relevant for the case of inferential robustness. A failure of completeness can, hence, be interpreted in two ways. Either one knows that none of the auxiliary assumptions supporting the inference is true or that one of the assumptions might be true. In both cases, it is not clear what the epistemic import of robustness analysis is. Naturally, the situation seems to be even less clear if both the completeness and the robustness condition fail to hold.

## 1.4 Taking stock and the main claims of the thesis

Let me take stock of the discussion so far. First, I revisit the taxonomy of robustness analysis put forward by Woodward (2006). Second, I revisit the arguments provided for the value of different types of robustness. As I go along, I flag the key questions that need to be addressed to assess the value of robustness analysis for evidence-based policy-making and state the responses that I will defend in the rest of this thesis.

As I have argued in the discussion of Woodward's taxonomy, inferential robustness analysis can be regarded as a special case of derivational robustness analysis. Hence, moving forward, I focus on derivational and measurement robustness analysis. The introductory examples of evidence-based policy-making suggested that both of these types of robustness analysis are prima facie relevant for policy-making. I will argue that derivational and measurement robustness analysis, in fact, play prominent roles in climate as well as economic policy-making. At this point, a natural question is the following:

**First research question** Do measurement and derivational robustness analysis exhaust the set of useful types of robustness analysis?

Eronen (2015, p. 3965) and Wimsatt (1981, p. 139, 144) view measurement and derivational robustness analysis as the fundamental types of this inference technique. I will argue that in relation to policy-making a further category of robustness consideration is useful: the stability of model results across different (types of) target systems. A model result is robust in this sense if it is instantiated across a number of target systems that differ substantially in their composition. I am going to call this type of robustness *predictive stability*. I show that toy models, on which economic policy-making relies, display predictive stability. This stability requirement can also be expressed in the form of an inference scheme. Let $i$ be an index defined over a set of $n$ target systems ($i = 1, ..., k, ..., n$) that differ substantially in their composition and $P(k)$ be the statement that the model result $P$ holds for target system $k$, then predictive stability expresses the following inference:

$$\frac{M \Rightarrow P}{\forall i : P(i)}$$

Let me now turn to the assessment of the value of robustness analysis. From the point of view of using robustness reasoning in policy-making, there is one pressing question: Are derivational and measurement robustness analysis sound inference techniques?

Let me start with derivational robustness analysis. The literature review presented two arguments for the confirmatory power of this type of robust-

ness. The first argument goes roughly like this: If robustness analysis shows that model results are driven by substantial model assumptions about causal factors or mechanisms, and there is evidence for these causal factors or mechanisms, then these results can be regarded as confirmed. The second argument goes roughly like this: If model results are robust across a set of models that accurately represent the space of relevant possibilities, then these results can be regarded as confirmed.

In my view, both of these arguments provide a sound foundation for the value of robustness analysis, that is, both inference patterns can convey confirmation on hypotheses given that the conditions laid out in the inference patterns are met. However, the two arguments describe derivational robustness analysis' role in confirmation differently. In the first argument, derivational robustness plays an indirect role in confirmation. According to this argument, robustness considerations can establish the fact that a model result is driven by substantial assumptions about the causal structure of a system. The confirmation is then provided by model-independent evidence on the correspondence of the causal structure postulated by the model and the one in the target system. Whereas, in the second argument, derivational robustness plays a more direct role in confirmation. Given that a set of models covers the relevant possibility space and the model results are in agreement, then the model result of interest is confirmed. In my view, the second argument is more relevant for evidence-based policy-making. For policy-makers often lack precise information about the causal structure of a target system. Consider again the case of the UK housing market. One problem of designing effective housing policy is precisely that it is unclear what factors determine the effectiveness of new regulation or subsidies.

However, with respect to this second argument, the review of the literature revealed two questions:

**Second research question** What is the value of derivational robustness analysis if not the entire relevant model space is covered by the available models?

I will argue that derivational robustness analysis can still have value if less

than the entire relevant model space is covered. Its value consists in the fact that it provides the resources to assess in a structured way how much of the relevant possibility space is covered by a model ensemble. This fact can be used to formulate preference relations over automated evidence aggregation procedures in medical policy-making. The other question that emerged from the literature review is the following:

**Third research question** What should one do if the model results are non-robust, that is, the model derivations are not in agreement?

I will argue that one must distinguish between problematic and unproblematic forms of a lack of derivational robustness where problematic cases are those that alter the lesson that must be drawn from a model substantially. With respect to problematic cases of a lack of derivational robustness, I show that one still can gain a form of conceptual learning or an uncertainty indication about values of variables in a target system. However, I will claim that these two types of insights are only of limited use for economic policy-making.

Let me now turn to measurement robustness analysis. The literature review has shown that an inference to the best explanation backs up the confirmatory import of measurement robustness: It would be a miracle if different ways of determining the truth of a hypothesis are all in agreement if the hypothesis were not true. The review also revealed that this no-miracle argument depends on a crucial assumption: the assumption that the different modes of determination are in fact independent from each other to a sufficient degree. Articulating this independence condition in a technically precise way has proven difficult. However, even if one does not attempt to provide a probabilistic characterisation of this independence condition, looking ahead to policy applications of measurement robustness analysis clearly presupposes that one has at least some practical, tangible guidelines to assess the independence of modes of evidence.[20] Going back to the initial example of medical

[20]In my view, it is the context of these concrete applications of measurement robustness analysis in which one can address Stegenga's complaint that the individuation of evidential modes is problematic. As will become clear across the case studies discussed in this thesis, in practical applications the different evidential modes can be straightforwardly individuated.

policy in the UK and my comments about evidence-based medicine reveals that in many policy contexts one particular mode of evidence, that is, expert opinion, plays a crucial role. The voice of prominent medical experts was a major factor behind the ongoing practice of bloodletting until the 19th century and the evidence-based medicine movement was introduced against the backdrop of what was called "expert based medicine" (Smith and Rennie 2014, p. 2). It appears that expert knowledge is often a key ingredient in the evidence mix that underlies policy decisions. Hence, one key question that emerges is the following:

**Fourth research question** How should one conceptualise the relation between expert knowledge and the evidence basis when one applies the framework of measurement robustness?

I shall argue that expert knowledge can only under special circumstances be viewed as a separate evidential mode that stands on a par with other evidential modes (such as observations or model outputs). For the main part, however, expert knowledge should not be viewed as separate to evidential considerations but rather as a type of knowledge necessary to address questions that arise when one applies the measurement robustness framework; questions such as whether evidential modes are independent and whether they are of high quality.

As with respect to derivational robustness analysis, the literature survey has made clear that measurement robustness can fail to hold. A natural reaction to this epistemic predicament is to suspend one's judgement about the hypotheses for which no robust evidence can be brought forward. However, particularly in policy-making contexts, one cannot always suspend one's judgement until the evidence is in agreement. It might simply be too costly to gather additional evidence that could resolve the disagreement or the time constraint might be such that one needs to decide based on a non-robust evidence set. Hence, the following question emerges:

**Fifth research question** What should one do if one faces a situation of non-robust evidence from multiple evidential sources?

I will argue that, in such a situation, selecting a theory or a model that underlie or constitute parts of the evidence set can be a reasonable action and I provide a procedure to make this choice in a rationally defensible way. This procedure suggests a way of seeing theory or model choices in a cardinal context, that is, a context in which information is available in the form of equally spaced units along a scale without a predefined zero point. Once this cardinal context is established, one can apply scoring rules, a tool from social choice theory, to make a choice. Scoring rules allow the aggregation of information from multiple criteria that are deemed relevant for the choice problem.

To summarise, here are my five research questions:

1. Do measurement and derivational robustness analysis exhaust the set of useful types of robustness analysis?

2. What is the value of derivational robustness analysis if not the entire relevant model space is covered by the available models?

3. What should one do if the model results are non-robust, that is, the model derivations are not in agreement?

4. How should one conceptualise the relation between expert knowledge and the evidence basis when one applies the framework of measurement robustness?

5. What should one do if one faces a situation of non-robust evidence from multiple evidential sources?

The chapter plan in the following section lays out how and in what order I approach these research questions. As will become clear, I do not address these five research questions in the order I have introduced them. This is because I have organised the chapters to reflect the policy domains of climate, medical and economic policy-making. I opted for this structure because, in my view, the questions do not suggest an order in which they should be addressed and the focus on policy domains emphasises the practical orientation of my inquiry.

## 1.5 Thesis outline

In *Chapter 2*, I look at the case of climate policy-making to address the fourth research question: How should one conceptualise the relation between expert knowledge and the evidence basis when one applies the framework of measurement robustness? In its most recent assessment report, the IPCC presented an updated version of the uncertainty framework. This framework is supposed to provide a unified approach to assessing and communicating uncertainties about predictions of climate variables, explanations of climate phenomena, as well as adaption and mitigation scenarios. The notion of a body of evidence, which is robust in a measurement robustness sense, figures prominently in this uncertainty framework. In this chapter, I analyse this uncertainty framework and argue that even a charitable interpretation of it faces substantial problems. These substantial problems allow me to delineate the relation between expert knowledge and the evidence basis in the context of a particular policy question. I argue that expert knowledge can only under special circumstances be viewed as a separate evidential mode that stands on a par with other evidential modes (such as observations or model outputs). Generally, however, expert knowledge should not be viewed as separate from evidential considerations.

In *Chapter 3*, I look at the case of medical policy-making to address the second research question: What is the value of derivational robustness analysis if not the entire relevant model space is covered by the available models? Recently, there has been a growing number of automated evidence aggregation procedures. My leading case study is Hunter and Williams (2012)'s proposal of a fully automated algorithm that inputs medical studies and yields treatment suggestions for a specific patient class. I argue that it is not clear how such automated evidence aggregation procedures should be assessed and I suggest a new way of conducting such evaluation. In particular, I argue that derivational robustness analysis can be a fruitful tool to determine the optimal extent of automation. Derivational robustness analysis turns out to be a structured way of establishing a preference relation over different algorithm designs by focussing on the amount of relevant possibility space. Due to the

analogy between algorithms and models, this analysis shows that derivational robustness analysis provides a structured way of forming preference relations over different model structures.

In Chapters 4 and 5, I look at the case of economic policy-making. In *Chapter 4*, I address the first research question: Do measurement and derivational robustness analysis exhaust the set of useful types of robustness analysis? One aspect, certainly not the only one, of economic policy-making is its reliance on theoretical models. In fact, a significant part of economic policy-making uses theoretical models that are extremely idealised and simple. These models, often called toy models, have recently received growing attention in the philosophy of science literature. I provide a new characterisation of this model class that consists of a manipulability, multiple realisability and a hybrid-representation condition. The multiple realisability condition expresses a new form of robustness, that is, the stability of model results across different (types of) target systems.

In *Chapter 5*, I address the third research question: What should one do if the model results are non-robust, that is, the model derivations are not in agreement? I use my characterisation of toy models to show that these type of models provide two distinct forms of non-explanatory learning: conceptual learning *from* a toy model and descriptive learning *with* a toy model. Having a clear grasp of these two types of learning allows me to argue that even non-robust model results can have epistemic value.

In *Chapter 6*, I address the fifth research question: What should one do if one faces a situation of non-robust evidence from multiple evidential sources? Given that a set of evidence can fail to be robust in a measurement robustness sense, one might need to choose a model or a theory that underlies the set of evidence. This brings into focus the question of how one should choose between theories and models. Recently, Okasha (2011) argued that such theory (or model) choice faces a predicament: Arrow's impossibility result. Arrow's impossibility result is the most famous result in social choice theory and can be put, roughly as follows: Given a set of plausible assumptions, there exists no aggregation function that maps individual preferences to a collective preference that is a weak ordering. Okasha (2011) argued that Arrow's im-

possibility result is relevant for the problem of theory choice: How should one evaluate a set of competing theories (or models and hypotheses) in light of epistemic virtues such as simplicity, accuracy, scope, fruitfulness, or consistency? In this chapter, I propose a solution to Okasha (2011)'s challenge. In a nutshell, the solution consists in offering a new way of thinking about how the problem of theory choice can be represented. If one has a cardinal informational structure, then Gaertner and Xu (2012)'s general scoring function can be used to aggregate the information.

# Chapter 2

# Robustness in Climate Policy-Making: An Assessment of the Uncertainty Framework of the IPCC

## 2.1 Introduction

At the Paris Climate Change Conference in November 2015, an agreement was reached that is regarded as a milestone in addressing global climate change. The parties committed to limit the increase in the global average temperature to below two degrees Celsius above the pre-industrial level (UN 2015, Article 2). In light of the high stakes, it can be hoped that the agreement is based on the best available scientific knowledge on the causes and consequences of climate change. Indeed, in the Paris Agreement an effective and progressive response to climate change is explicitly asked for "in light of the best available scientific knowledge" (UN 2015, Preamble).

The Intergovernmental Panel on Climate Change (IPCC) synthesizes the current state of knowledge on climate change (Mastrandrea et al. 2011, p. 676). This involves understanding and being able to communicate the uncertainties surrounding these scientific findings. To achieve these ends, the IPCC

has developed an uncertainty framework (Mastrandrea et al. 2011, p. 676). This framework is best understood as an attempt at a unified conceptualisation of different types of uncertainties, including model and data uncertainty, scenario uncertainty, as well as ethical uncertainty (IPCC 2013, p. 138; Mastrandrea et al. 2011, p. 676).[1] The framework is in constant development with later versions building on previous ones. The latest and most developed version of the framework equips scientists with a confidence and likelihood metric to qualify their statements.

Within this framework the notion of robust evidence plays a crucial role. What notion of robustness is in play here? The uncertainty framework defines robust evidence as "multiple, consistent, independent lines of high-quality evidence" (Mastrandrea et al. 2011, p. 678). This notion is best understood as measurement robustness. To see this, let me point out two aspects of robust climate scientific evidence. To start, for most of the questions that the IPCC assesses, one faces a situation of evidential diversity: evidence from multiple sources such as current or historical observational data, model outputs, and theories is available and synthesized. For example, the existence of marine-ice sheet instability is assessed by taking ice-dynamics theory, numerical modelling simulations, and paleo records into account (IPCC 2013, p. 1174). Furthermore, the IPCC's definition of robust evidence highlights conditions similar to the ones encountered in the introduction, which need to be in place such that concurrence or stability of results across different types of evidence has special epistemic merit. Interestingly, one of the key claims of the framework is that robust evidence in combination with high agreement amongst experts confers very high confidence onto climate scientific findings (Mastrandrea et al. 2010, p. 2). As I detail below, the IPCC provides an elaborate matrix on how expert agreement and evidence should influence confidence judgements on scientific findings.

Given this, the uncertainty framework proves to be an ideal case study for investigating my fourth research question: How should one conceptualise the relation between expert knowledge and the evidence basis when one applies the framework of measurement robustness? I am going to argue that ex-

[1] I introduce these different types of uncertainty below.

pert knowledge can only under special circumstances be viewed as a separate evidential mode that stands on a par with other evidential modes (such as observations or model outputs). For the main part, however, expert knowledge should not be viewed as separate to evidential considerations but rather as a necessary type of knowledge to address questions that arise when one applies the measurement robustness framework; questions such as whether evidential modes are independent and whether they are of high quality.

I defend this claim by pointing out the conceptual shortcomings of the current framework. A significant body of literature exists that discusses previous and current versions of the IPCC's uncertainty framework (see Adler and Hadorn 2014 for a review). The majority of the literature focuses on issues surrounding the interpretation of probabilistic information by the readers of the IPCC reports (e.g., Budescu et al. 2014; Morgan 2014). However, the literature does address the conceptual foundations of the latest version of the uncertainty framework only in a limited way (see Aven and Renn 2015; Jones 2011; Socolow 2011).

This chapter is structured as follows. First, I give some background on the structure and aim of the IPCC and the motivation that lies behind the uncertainty framework (*Section 2.2*). I then put forward an interpretation of the framework that draws from introductory documents and the actual practice of its users (*Section 2.3*). With this interpretation in hand, I identify three conceptual problems that point towards untenable assumptions regarding evidence aggregation in the context of climate scientific findings (*Section 2.4*). This analysis allows me to defend the main claim of this chapter (*Section 2.5*). I close by putting forward three tentative suggestions for improving the uncertainty framework (*Section 2.6*).

## 2.2 Background: Structure and aim of the IPCC

The IPCC was established by the United Nations Environment Programme and the World Meteorological Organization in 1988 (IPCC 2014a, p. 1). The

aim of the IPCC is to "provide the world with a clear scientific view on the current state of knowledge in climate change and its potential environmental and socio-economic impacts" (IPCC 2014a, p. 1).

The IPCC does not conduct original research but synthesises current scientific knowledge on climate change. Furthermore, the IPCC does not provide policy recommendations: "the work of the organization is (...) policy-relevant and yet policy-neutral, never policy prescriptive." (IPCC 2014a, p. 1) Hence, the IPCC can be understood as an aggregation body of scientific evidence that supplies policy-makers with a crucial but not sufficient component for their decision making.

The key instrument for the interaction of the IPCC with policy-makers is the assessment report. The latest, fifth assessment report was published in 2014. The assessment report consists of three parts written by three separate working groups (IPCC 2014b, p. 1). Working group I is concerned with the assessment of the physical aspects of the climate system and its change. Working group II assesses the vulnerability of socio-economic and natural systems to climate change and processes of adaptation. Working group III reviews the options for mitigating climate change based on limiting greenhouse gas emissions and activities to remove them. For every working group report there exists a summary for policy-makers, which states the key points of the analysis in a non-technical language. A synthesis report across all three working groups is also provided (IPCC 2014b, p. 1).

The work within these groups is organised roughly as follows. Coordinating lead authors are assigned to chapters of the respective assessment report. These authors coordinate the work of lead authors and contributing authors. The resulting document is reviewed by review editors. For working group I, 209 coordinating lead authors and lead authors as well as 50 review editors and more than 600 contributing authors were part of the process. I mention these details to convey the large amount of coordination that is involved in producing an assessment report. In addition, summaries for policy-makers are subjected to one round of governmental comments (IPCC 2013, p. viii). This process of governmental approval highlights the political nature of the IPCC assessment report.

Working group I distinguishes between four types of uncertainties: first, *scenario uncertainty*, referring to uncertainties due to limited understanding regarding future emissions, concentrations and forcing trajectories, and lack of knowledge about future options for adapting and mitigating climate change, such as carbon capture methods (IPCC 2013, p. 138); second, *model uncertainty*, capturing uncertainty regarding model parameters and the causal structure of the system under study (IPCC 2013, p. 138); third, *internal variability uncertainty*, indicating uncertainties involved in describing the variability of earth's climate in absence of any forcing due to emissions (IPCC 2013, p. 138); fourth, *boundary condition uncertainty*, denoting uncertainties regarding historical and paleoclimate simulations (IPCC 2013, p. 139). The last two types of uncertainty are subsumed under the heading of data uncertainty. *Ethical uncertainty* is the lack of knowledge regarding how outcomes should be evaluated. Here the problem is not a lack of empirical knowledge (i.e., one knows what is going to happen given a certain action) but rather a lack of normative knowledge (i.e., how one should value the particular outcome) (Bradley and Drechsler 2014, p. 1228).

## 2.3 The uncertainty framework of the IPCC: An interpretation

To adequately account for the uncertainties involved in scientific findings, the IPCC uses in its latest assessment report an updated version of the uncertainty framework.[2] This framework serves two functions. It is an analytical instrument to understand uncertainty and a tool for communicating uncertainties to policy-makers (Mastrandrea et al. 2010, p. 1).

[2]A comparison of the current version of the uncertainty framework, published in 2010 and used in the Fifth Assessment Report, with the previous one, published in 2005 (and used in the Fourth Assessment Report published 2007), can be found in Mastrandrea et al. (2010, Annex A). Here, I engage with the latest version of the framework for two reasons: First, each version of the framework builds on earlier ones and takes shortcomings into account; hence, it makes sense to engage with the most recent, arguably, most refined, framework. Second, solely the latest version of the uncertainty framework is in use currently and is, therefore, policy-informing.

Throughout this chapter, I will focus on the assessment report of working group I.[3] The IPCC provides two supporting documents which explain the framework: the Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties (Mastrandrea et al. 2010) and a commentary article by Mastrandrea et al. (2011). I also take into account the practice of the authors of the assessment report. As the discussion below will make clear, the supporting documents and the practice reveal ambiguities and inconsistencies in the framework. Hence, the aim of this section consists in providing a charitable interpretation of the framework, which dissolves the ambiguities in a way that minimises inconsistencies.

To gain a handle on the uncertainty framework, consider the following examples of its application. To start, there are cases in which solely a confidence term is used to characterize a finding:

> The release of CO2 or CH4 to the atmosphere from thawing permafrost carbon stocks over the 21st century is assessed to be in the range of 50 to 250 GtC [giga tons of carbon] (...) (*low confidence*). (IPCC 2013, p. 27, my emphasis)

There are cases in which solely a likelihood term is used:

> It is *likely* that the frequency of heat waves has increased in large parts of Europe, Asia and Australia [since the 1950s]. (IPCC 2013, p. 5, my emphasis)

There are cases in which both confidence and likelihood terms are used:

> In the Northern Hemisphere, 1983-2012 was *likely* the warmest 30-year period of the last 1400 years (*medium confidence*). (IPCC 2013, p. 3, my emphasis)

Finally, there are also cases in which no qualifying term is used. The absence of qualifying terms indicates a finding for which "evidence and understanding are overwhelming" (Mastrandrea et al. 2010, p. 2).

[3] I will briefly address the question of how my insights generalise to reports of working groups II and III in the concluding remarks of this chapter.

How should these confidence and likelihood qualifications be understood and how are they related to each other?

### 2.3.1 Likelihood terms

Let me begin with the likelihood terms. The guidance note reveals that these terms correspond to ranges of probabilities: virtually certain (probability of the occurrence of the outcome is 99-100%), very likely (90-100%), likely (66-100%), about as likely as not (33-66%), unlikely (0-33%), very unlikely (0-10%), and exceptionally unlikely (0-1%) (Mastrandrea et al. 2010, p. 3). Accordingly, these likelihood terms express a quantitative measure of uncertainty (Mastrandrea et al. 2010, p. 1).

The basis for the ascription of probabilistic information can be statistical, modelling analysis or elicitation of expert views (Mastrandrea et al. 2010, p. 3). Model analysis involves the analysis of time series data for a variable (e.g., global mean surface temperature) over a period of interest (e.g., 1980-2050) from single or multiple models. It is with respect to this analysis that derivational robustness analysis comes into play. For example, an ensemble of models is used to generate predictions about the sea surface temperature over the eastern equatorial Pacific Ocean and the variability of the predictions is assessed (IPCC 2013, p. 107). Expert elicitation techniques are procedures that aim to determine experts' subjective degrees of belief about the value of variables. They are primarily used to capture the meta-knowledge of experts regarding the limitations of climate models and observational data (Morgan 2014, p. 7176).

### 2.3.2 Confidence terms

Let me now turn to the confidence terms. In contrast to probabilistic information, confidence is expressed qualitatively, that is, it can be very high, high, medium, low, or very low (Mastrandrea et al. 2010, p. 1). The guidance note gives the following indication for arriving at these confidence statements (Mastrandrea et al. 2010, p. 1, my emphasis):

> Confidence in the validity of a finding. Based on the type, amount, quality, and consistency of *evidence* (e.g., mechanistic understanding, theory, data, models, expert judgement) and the degree of *agreement*. Confidence is expressed qualitatively.

This statement needs to be disentangled. I begin with the two notions of evidence and agreement. I then shed light on how these two notions are combined to arrive at confidence statements. I pay particular attention to the questions of how expert judgement and expert agreement are treated.

The supporting documents suggest an assessment of the available evidence on the basis of the type, amount, quality, and consistency of evidence. These four dimensions are best viewed as criteria that should be considered individually by the authors of the assessment report (Mastrandrea et al. 2010, p. 2).

Under the heading of *type of evidence*, the guidance note provides five categories of evidence: mechanistic understanding, theory, data, models, and expert judgement (Mastrandrea et al. 2010, p. 1). This underscores the point I made earlier about the evidential diversity that is considered relevant by the authors. Mechanistic understanding is described as understanding of the physical processes governing a particular phenomenon (Mastrandrea et al. 2011, p. 678). As Jones (2011, p. 737) notes, it is puzzling why theory is treated as one category of evidence, since, normally, evidence is viewed to be (dis)confirming theories. A charitable way of understanding theory as a subcategory of evidence is to see it as theoretical knowledge supporting (or undermining) the particular explanations or predictions that are reported in a finding. Predictions can be supported by theory, for example, if the predictions are model-based (and the theory supports the structural assumptions of the model) or are based on expert elicitation (and the experts ground their judgements in theory). This interpretation of theory as an evidence category is suggested by the practice of the authors:

> In summary, *ice-dynamics theory*, numerical simulation, and paleo records indicate that the existence of a marine-ice sheet instability

(...) is possible in response to climate forcing. (IPCC 2013, p. 1174, my emphasis)

*Amount* and *quality* of evidence are not defined in the supporting documents. The authors use amount of evidence to denote different things: a) number of observation points (e.g., IPCC 2013, p. 137, 158); b) number of models or total number of scenarios run on selected models (e.g., simulation of Greenland ice sheet, Figure 5.16, IPCC 2013, p. 428); or c) the number of studies (e.g., IPCC 2013, p. 129). Quality of evidence is used by the authors in relation to observational data and models. Data quality involves judgements about instrument design, equipment handling, or data processing (e.g., IPCC 2013, p. 143). Model quality is assessed based on empirical model performance and adequate representation of relevant causal factors (IPCC 2013, p. 749, 753f.).

*Consistency* of evidence is defined as "(...) the extent to which it [evidence] supports single or competing explanations of the same phenomena, or the extent to which projected future outcomes are similar or divergent." (Mastrandrea et al. 2011, p. 678)

Evidence is expressed on a qualitative scale: evidence can be robust, medium, or limited (Mastrandrea et al. 2010, p. 2). As already stated in the introduction, robust evidence should be understood in the sense of measurement robustness and is defined as multiple, consistent independent lines of high-quality evidence (Mastrandrea et al. 2011, p. 678). Notice that here an additional criterion for the evaluation of evidence enters the scene that is not explicitly introduced in the framework: the (in)dependence of different pieces of evidence. For example, the models in an ensemble can be independent to a higher or lower degree, given how many model assumptions they share (IPCC 2013, p. 755). The supporting documents do not define the levels of medium and limited evidence. In particular, there are no aggregation rules given that might indicate the relative importance of type, consistency, independence, amount, and quality of evidence. The practice of the authors does not reveal specific aggregation rules or principles either.

Let me now turn to the notion of *agreement*. Agreement is expressed qual-

itatively: agreement can be high, medium, or low (Mastrandrea et al. 2010, p. 3). Agreement is not defined in the guidance note. However, Mastrandrea et al. (2011, p. 678) offer the following two accounts of agreement in their commentary:

> [Agreement] is the level of consensus in the scientific community on a particular finding.

> [Agreement indicates] the degree to which a finding follows from established, competing, or speculative scientific explanations.

At first sight, these two definitions differ. A way of reconciling them would be to add the assumption that the level of consensus in a scientific community depends on the degree to which a finding follows from established, competing, or speculative scientific explanations. Given this assumption, the second definition entails the first one.

Interpreting the framework in the way suggested here leads to a problem: agreement and consistency cannot be ascribed independently from each other. Recall that both agreement and consistency are defined with respect to a finding in the assessment report. However, if agreement is low (due to the presence of competing explanations), then consistency needs to be low as well, since consistency is defined as the number of explanations supported by the evidence. The same holds for high agreement and high consistency. Mastrandrea et al. (2011, p. 678) seem to sense this tension and explain the difference between agreement and consistency as follows:

> "Agreement is not equivalent to consistency. Whether or not consistent evidence corresponds to a high degree of agreement is determined by other aspects of evidence such as its amount and quality; evidence can be consistent yet low in quality."

This explanation is not satisfactory since it is in tension with our first attempt to make sense of agreement. Agreement has been so far understood as depending only on the number of competing explanations for a finding. Here, Mastrandrea et al. (2011) suggest that it depends also on the amount

and quality of evidence. The best way to circumvent this problem is to view agreement as an umbrella notion that covers two different concepts: agreement as degree of consensus in the scientific community and agreement as consistency of evidence.

This interpretational move is supported by the practice of the authors. When the authors are using the uncertainty framework, they interpret agreement in the majority of cases as consistency of evidence. Let me give an example:

> *High agreement among analyses* provides medium confidence that oxygen concentrations have decreased in the open ocean thermocline in many ocean regions since the 1960s. (IPCC 2013, p. 52, my emphasis)

However, there are also instances where agreement is viewed as consensus in the scientific community:

> Many semi-empirical model projections of global mean sea level rise are higher than process-based model projections (...), but there is no consensus in the scientific community about their reliability and there is thus low confidence in their projections. (IPCC 2013, p. 26)

Let me take stock at this point. A hierarchy of the notions introduced through the uncertainty framework, given the interpretational ambiguities, can be visualised (see Figure 2.1).

After having discussed the evidence and agreement notions, I now turn to their *aggregation into overall confidence statements*. The supporting documents specify that the increase in levels of agreement or evidence (individually, while holding the other constant, or together) leads to a rise in the confidence level (Mastrandrea et al. 2010, p. 3). For findings with high agreement and robust evidence, the confidence level 'very high' should be assigned (Mastrandrea et al. 2010, p. 2). For findings with either high agreement or robust evidence a confidence level should be given if possible (i.e., high confidence or

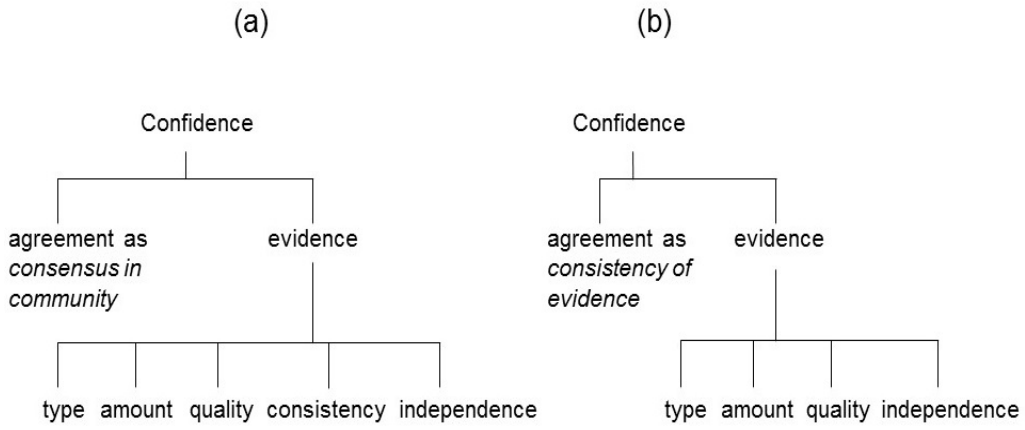


Figure 2.1: Elements of the confidence metric under two interpretations of agreement: (a) agreement as consensus in the scientific community, (b) agreement as consistency of evidence (my diagram).

medium confidence). If this is not possible, then the summary terms should be used (e.g., robust evidence, medium agreement) (Mastrandrea et al. 2010, p. 3). For findings with low agreement and limited evidence, the summary terms should be used. Figure 2.2 visualises these rules.

So far, I have discussed the likelihood and confidence terms of the uncertainty framework. A crucial question remains: How are these metrics related to each other?

### 2.3.3 The relation of likelihood and confidence terms

Do likelihood and confidence terms convey the same or different types of information? The supporting documents do not rule out either of the two possible but incompatible answers.

To start, the likelihood metric can be interpreted as a quantified uncertainty tool that co-varies with the confidence metric. This interpretation would treat the likelihood and confidence metrics as substitutes, conveying

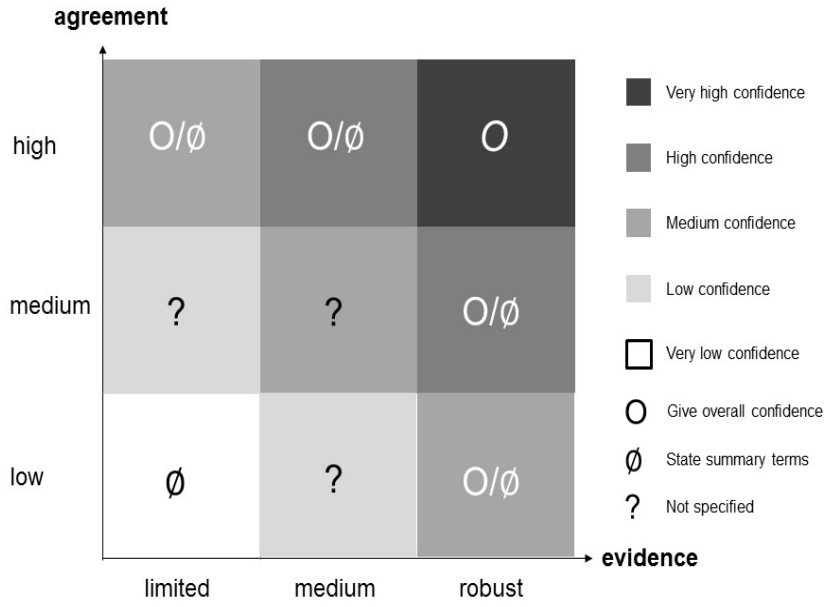


Figure 2.2: Aggregation of evidence and agreement into overall confidence statements (my diagram, based on Mastrandrea et al. 2010, p. 2-3).

the same information. Any difference in the application of the two metrics would derive from the fact that not all types of evidence allow a quantified treatment of uncertainty. I refer to this reading as the *substitutional interpretation*. The following statements from the supporting documents back this reading:

> Depending on the nature of the evidence evaluated, teams have the option to quantify the uncertainty in the finding probabilistically. (Mastrandrea et al. 2010, p. 1)

> [If] a range can be given for a variable, based on quantitative analysis or expert judgement: Assign likelihood or probability for that range when possible; *otherwise* only assign confidence. (Mastrandrea et al. 2010, p. 4, my emphasis)

However, one can also interpret the relation between confidence and likelihood statements differently: Confidence statements can be viewed as meta-judgements and likelihood statements as intra-finding judgements. According to this interpretation, the confidence metric allows assessing the goodness of the evidential basis of a finding whereas the likelihood metric can be used to specify the events that are mentioned in that finding. Under this reading, the two metrics would convey different information. This would allow the absence of co-variation between the likelihood and the confidence metric. I refer to this reading as the *non-substitutional interpretation*. The following statements from the supporting documents back this interpretation:

> Author teams are instructed to make this evaluation of evidence and agreement the basis for any key finding, *even those that employ other calibrated language* (...). (Mastrandrea et al. 2010, Annex A, my emphasis)

> This scale [confidence metric] may be *supplemented* by a quantitative probability scale (...). (Mastrandrea et al. 2010, Annex B, my emphasis)

How do the authors of the assessment reports deal with this interpretational ambiguity? The practice reveals that in the clear majority of cases authors opt for the non-substitutional interpretation. The following example illustrates this:

> Estimates of the Equilibrium Climate Sensitivity (ECS) based on multiple and partly independent lines of evidence (...) indicate that there is *high confidence* that ECS is *extremely unlikely* to be less than 1°C (...). (IPCC 2013, p. 871, my emphasis)

Finally, the supporting documents prescribe the use probabilistic information only if confidence in a finding is high or very high (Mastrandrea et al. 2010, p. 4). No rationale is given for this rule. The implicit motivation could be that it is more problematic to assign probabilities given one has low confidence in the evidential basis for a finding. The authors of the assessment

report seem to disregard this rule about the use of probabilistic information. There are multiple instances in which likelihood terms are used given very low, low, or medium confidence. Here is an example:

> A nearly ice-free Arctic Ocean (...) in September before mid-century is likely under [emission scenario] RCP 8.5 (medium confidence). (IPCC 2013, p. 92)

A straightforward way of reconciling the practice of the authors with the guidance note would be to interpret the rule as stating that precise probabilistic information, that is, complete probability density functions, should only be given if confidence is high or very high.

This concludes my attempt at giving a coherent interpretation of the uncertainty framework. In the next section, this interpretation will serve as a background to an engagement with the conceptual foundations of the framework.

## 2.4 Conceptual problems in the fundament of the uncertainty framework

In this section, I argue that the uncertainty framework exhibits three substantial conceptual problems. These problems make clear that the current version of the framework is neither an adequate tool for conceptualising the uncertainties involved in climate scientific findings nor for communicating them to policy-makers.[4] Importantly, identifying these three conceptual problems also paves the way towards assessing the leading question of this chapter: how should expert judgement be treated in measurement robustness considerations? In Section 2.5, I argue that there are two distinct roles for expert judgement with respect to measurement robustness: First, expert judgement

[4]Aven and Renn (2015) and Jones (2011) highlight additional conceptual problems of the framework. For example, the problem of the consistent application of a qualitative (compared to quantitative) confidence metric, the ambiguity of likelihood as an estimated value or a belief about the true underlying likelihood, and the omission of the distinction between perceived and actual risk. These observations are complementary to my analysis.

can play a crucial role in assessing conditions that need to be in place such that measurement robustness delivers its epistemic goods, and, second, under some specific conditions, expert judgement can be a source of evidence. To establish these claims, particular attention needs to be paid to the role of *expert agreement* to which I turn now.

The first problem concerns the *bifurcation of evidence and agreement in the confidence metric*. Given the two possible interpretations of agreement, this bifurcation does not hod up to scrutiny. If one understands agreement as consensus in the scientific community, then the social fact of consensus should be a result of the evidence and should not be treated as an independent dimension. As thermometer readings should track temperature, the social fact of consensus in a community should supervene on the available evidence. Ultimately, it is the evidence that should guide our uncertainty assessment. If one understands agreement as consistency of evidence, then agreement is straightforwardly part of the evidence dimension and nothing separate from it.

This bifurcation between evidence and agreement leads to a second problem unfolding into a set of issues related to the *rules for aggregating evidence and agreement statements*.

To start, as illustrated in Figure 2.2, the uncertainty framework allows for a combination of robust evidence and low or medium agreement. Recall that 'robust evidence' is defined as multiple, consistent independent lines of high-quality evidence (Mastrandrea et al. 2011, p. 678). If one understands agreement as consistency, then there cannot be low (or medium) agreement in the light of robust evidence. Since robust evidence involves evidence that is consistent, agreement as consistency needs to be high in the light of robust evidence. If one understands agreement as consensus in the scientific community, then it is puzzling how there can be a limited level of consensus in the light of robust evidence, given one makes the minimal assumption that scientists base their judgements on the available evidence.

Even if one disregards this issue, a second issue emerges in relation to aggregation rules. The pairs 'limited evidence/high agreement' and 'robust evidence/low agreement' are treated symmetrically by assigning them medium

confidence. This symmetry is puzzling. The intuition that one faces less uncertainty given 'robust evidence/low agreement' than 'limited evidence/high agreement' seems natural. This intuition can be substantiated as follows. If one understands agreement as consensus in the scientific community, it seems questionable to give the fact of consensus the same weight as evidential considerations. If one understands agreement as consistency, it seems problematic to give consistency considerations the same weight as the combined considerations about type, amount, quality, and independence of evidence. This point gains traction in the practice of the authors of the assessment report. Instead of weighting consistency in the light of other evidential considerations, the authors solely use consistency considerations to arrive at overall confidence statements. Here is an example:

> *High agreement among analyses provides medium confidence* that oxygen concentrations have decreased in the open ocean thermocline in many ocean regions since the 1960s. (IPCC 2013, p. 52, my emphasis)

This constitutes bad epistemic practice since the other dimensions, which should enter a critical assessment of the underlying evidential basis for a finding, that is, the type, amount, quality, and independence of the evidence, are not considered. If one assumes that the authors are considering these alternative dimensions implicitly, this practice is not transparent to the readers of the report.

A final issue concerns the amount of information that is provided about the aggregation rules. Consider again Figure 2.2. Why is it the case that sometimes one can give an overall confidence judgement and sometimes one is only supposed to give the summary terms? It is conceptually unclear how a line between these two classes of cases can be drawn. Furthermore, the diagonal in Figure 2.2 is puzzling. The diagonal contains all matching evidence and agreement pairs (e.g., limited evidence, low agreement). In these clear-cut cases, an overall confidence statement should be possible. However, as Figure 2.2 illustrates, this is not the prescription of the framework.

The third problem can be located in the *rules about when it is permissible to use probabilistic information.* A charitable interpretation of the uncertainty framework is that only if confidence is high or very high, precise probabilistic information should be used to express uncertainty. If one adopts the non-substitutional reading as suggested by the practice of the authors, it is not clear why one cannot use precise probabilistic information if the confidence in the evidential basis yielding this probabilistic information is low. To start, using probabilities in itself does not confer any epistemic merit on a finding. Furthermore, to prohibit communicating probabilistic information in these cases amounts to deliberately setting aside available information. This violates Carnap's plausible principle of total evidence (Carnap 1947), and, hence, should alert suspicion.

Let me pause at this point and put these problems into the context of my overall argument in this chapter. The first two problems are directly related to my leading question of how expert judgement should be brought into contact with the machinery of measurement robustness analysis. Hence, these are the focus of my attention going forward. The upshot of these first two problems is that expert agreement cannot be viewed as a separate, additional consideration to the evidence that feeds into measurement robustness analysis. In the remainder of this section, I strengthen this claim by considering some potential objections.[5]

A *first objection* might go like this: Expert agreement understood as consensus in the scientific community must be treated as independent from evidence since the evidence itself does not settle the issue. To put this in Bayesian terms, different experts might approach the evidence with heterogeneous priors, and, hence, end up with different posteriors in the light of the same evidence.[6]

Let me make two points in response to this objection. First, it must be noted that Bayesian convergence theorems can be employed that show that

[5]By considering *potential* objections, I am able to include parts of the literature that so far have not been brought into direct contact with the uncertainty framework, but that are relevant from a systematic point of view.

[6]I thank an anonymous reviewer who commented on the shorter, published version of this chapter for raising this criticism.

agents with heterogeneous priors will converge on their posteriors, given a series of shared observations. I grant that these theorems come with specific conditions attached – without going into the technical details of the literature – such as at least some agreement in priors, numerous observations, or a long time span (see for example Doob 1971). However, it is not clear to me why these conditions cannot be met by the practice of climate scientists, especially by those groups of the scientific community who are working on particular topics (such as glacial hydrology). Second, one does not need to rely on convergence theorems to defuse this objection. Consider that the experts are supposed to reach an agreement on the evidence dimension separately, that is, they need to agree whether there is robust evidence. As a reminder, robust evidence is defined as multiple, consistent independent lines of high-quality evidence. I take it to be the case that agreement on the presence of multiple, consistent independent lines of high-quality evidence is hardly compatible with experts maintaining radically different posteriors, no matter how different their priors for a specific hypothesis were. Most importantly, the consistency requirement of robust evidence, forces a high degree of agreement about predictions or explanations that are deemed defensible in the light of this evidence. The key point to see is that evidence in the uncertainty framework not only denotes shared observations but involves a statement about the relationship between evidence and a hypothesis.

Socolow (2011, p. 785-786) raises a *second objection* by offering the following argument for the independence of evidence and agreement: Robust evidence can be combined with low agreement if key information is missing. If key information is missing, given a large set of high quality and consistent evidence, a variety of projections and explanations of a phenomenon are possible, which is the definition of low agreement in terms of consistency of evidence. Furthermore, limited evidence can be combined with high agreement if one is in a situation in which the experts have converged on a single explanation although robust evidence is lacking.

I think this line of reasoning is not compelling for two reasons. First, according to the uncertainty framework, the claim that robust evidence is available requires that this evidence is consistent, and, hence – given the defi-

nition of consistency –, the evidence does not support diverging explanations or predictions. Therefore, it is unclear how there can be competing explanations when experts face robust evidence. Second, the number of competing explanations for a finding is not conceptually linked to the agreement about the finding, since scientists could agree on key characteristics of a phenomenon even in the light of competing explanations for the occurrence of this finding.

A *third objection* might point to the source of expert agreement. Douglas (2012, p. 152) points out that expert disagreement about hypotheses in the light of shared evidence is grounded in different explanations about why the evidence appears as it does. Hence, expert disagreement is possible in relation to the same set of evidence.

Although this argument might be applicable to many areas of expert disagreement, the agreement (or disagreement) in the context of the IPCC's uncertainty framework does not fall within its scope. Consider again the fact that a working group must come to a shared judgement about the state of the evidence for a hypothesis. This judgement – limited, medium, or robust evidence – requires experts to take a stance on whether the evidence is consistent (i.e., whether it supports a wide or narrow set of predictions or competing explanations). Both the explanations and the predictions must be viewed in relation to the hypothesis in question. Hence, experts express judgements about whether they agree or disagree on a hypothesis in question. Any reason for this disagreement – which might well be of the sort that Douglas (2012) mentions – is not relevant input for the uncertainty framework.

A *fourth objection* runs as follows: The reason why one should value expert agreement independently of the evidence set is that expert agreement contains information about those components of expert judgement that cannot be expressed in terms of characteristics of the evidence. A case in point could be, arguably, the agreement of experts on a set of tuning parameters for a complex climate model. Tuning parameters are those parameters that can be changed in a fine-grained manner to increase the empirical fit of a climate model. It might be the case that experts are not able to justify the parameters in the light of the available evidence. The distinction between knowing that and knowing how can substantiate this point. If climate scientific research

requires a significant amount of abilities that cannot be articulated fully in evidential terms, then the uncertainty framework for climate scientific findings should account for them. If there is agreement among experts regarding the outcomes of these inarticulable abilities, then it makes sense to treat expert agreement as an independent dimension in relation to the evidence dimension.

Let me make two points in response to this claim. First, in my view the burden of proof lies with those who argue that there are significant parts of climate scientific practice that cannot be articulated in evidential terms. Before one can assess the strength of this claim, it is desirable to have a clear grasp of what areas of climate-scientific practice and what types of activities are involved. Second, even if this can be made more precise and it turns out that there are significant parts of climate-scientific practice that cannot be articulated in evidential terms, an uncertainty framework should be maximally transparent about these parts. Merely subsuming them under the heading of expert agreement does not communicate the necessary information.

A *fifth objection* is this: There are scenarios in which one does not have evidence for a hypothesis and, hence, one should base uncertainty assessments on the level of expert agreement in terms of a consensus in the scientific community (see Bradley et al. 2016, p. 9 for this claim).

In the hypothesised situation, evidence is absent. Accordingly, expert agreement is not a separate evaluative dimension compared to evidence but the only remaining one. Given this, it seems natural to view expert agreement as a proxy to the unavailable – potentially not yet gathered or not yet analysed – evidence. However, expert agreement is only a reliable proxy if certain conditions are met. For example, experts should not be biased individually (e.g., by reporting their assessment in a strategic manner) or as a group (e.g., by relying on group judgement aggregation mechanisms that do not adequately represent the judgements of group members). Furthermore, the experts should be chosen such that they have the relevant expertise to assess a hypothesis despite the fact that there is no evidence which bears directly on it. However, even if these conditions are met, a satisfactory uncertainty framework should contain the requirement that the experts are as transparent as possible about the reasons for their agreement (or lack of agreement).

Simply reporting the degree of agreement in situations of absent evidence is not transparent enough since these reasons are ultimately ground uncertainty assessments.

Following Reiss (2015), a *sixth objection* runs as follows: The IPCC provides risk assessments. These risk assessments are done by working groups II and III for mitigation and adaption scenarios. Risk assessments should be evidence-based. In the IPCC set-up, these risk assessments are based on the scenarios and the process understanding provided by working group I. Now, evidence assessments always involve norms that themselves cannot be judged against the facts, meaning that these questions cannot be settled by the evidence itself or other empirical facts but involve a non-reducible normative component. Normative decisions of this particular kind are, for example, decisions about the appropriate level of significance for rejecting a hypothesis (in a frequentist approach to statistical testing); or which concept of cause is the appropriate one for a particular context. Hence, the uncertainty framework should give the role of expert agreement an even larger role than it already does.

The strength of this argument depends on what context of risk assessments one examines. In the case of climate scientific risk assessments, the fact that the IPCC has established a review process that defines shared norms of assessment should reduce the likelihood that there are radically different methodological or conceptual standards in play that necessitate making expert agreement central. Furthermore, giving up the evidence dimension with its categories that could guide the process of uncertainty assessment is misguided since it seems to presuppose that these categories do not provide a fruitful framework for most of the cases covered by experts. However, let me emphasise once more that my claim is not that expert agreement should not play a role in uncertainty assessments. My criticism solely targets the claim that expert agreement understood as social consensus in the scientific community (and as consistency of evidence) is an independent dimension that grounds confidence judgements in a hypothesis.

Finally, a *seventh objection* can be mounted if one leaves the realm of the epistemic considerations and pays attention to the social context of the IPCC.

The IPPC supplies decision-relevant information to political leaders who face high-stakes decisions. These high-stakes decisions – such as cutting carbon emissions – have significant implications for the (immediate and long-term) well-being of their respective electorate. In this context, expert agreement can have social benefits. Expert agreement signals the coherence of scientific opinion and, thereby, can underwrite claims about the urgency of action regarding climate policy.

I do think that this social function of coherence is undeniable, however, I do not think that – even in the case of the IPCC – the blending of social and epistemic considerations is desirable. To start, it is not clear that the experts agree on the hypotheses and, hence, giving expert agreement the weight of an independent dimension of assessment can be counterproductive. Furthermore, the social benefits can still be realised without disregarding the need for sound epistemic foundations of the uncertainty assessment. They can be realised by being more aware of the division of cognitive labour between the policy-informing IPCC authors and the policy-conducting governing bodies behind the IPCC. The degree of certainty in scientific findings can be assessed based on the evidence at hand. Once the degree of certainty is assessed, the policy-making bodies can devise decision mechanisms that allows them to arrive at policies that, then, can be communicated.

## 2.5 The role of expert judgement in measurement robustness analysis

In the previous section, I put forward the claim that expert agreement and evidence should not be viewed as two independent dimensions of assessment in uncertainty judgements. A natural follow-up question is to ask what the role of expert judgement (*not* the fact that experts agree) should play in uncertainty assessments. Since uncertainty assessments in the context of the IPCC should be based on evidence, and the IPCC is facing a situation of evidential diversity, this question boils down to asking how expert judgement should be related to measurement robustness considerations. In my view, the

discussion so far points towards two distinct roles for expert judgement with respect to measurement robustness.

To start, expert judgement can play a role in assessing conditions that need to be in place so that measurement robustness delivers its epistemic goods. As the discussion of the framework has shown, these conditions refer to the quality of evidence, the number of evidence pieces, and their independence. Assessing these three dimensions of evidence requires, plausibly, some expert judgement. It is not the case, certainly not for the context of climate scientific findings, that these dimensions are defined rigidly and are unanimously agreed upon. To see this, consider why the IPCC developed the uncertainty framework in the first place; it was introduced to provide a shared standard of evidence appraisal. Also, the practice of the authors in working group I suggests that arriving at statements about the nature of the pool of relevant evidence are the very subject of the discussion in the author teams compiling the respective sections of the report. It is a further interesting question how experts should express their individual statements about quality, amount, and independence and how these statements should be aggregated. I will not dive into this question here to avoid distracting from the main point of this section.[7]

Furthermore, expert judgement can be one type of evidence that feeds into measurement robustness analysis. This is the case if the elicited expert judgement concerns one (or multiple) aspects of a problem (e.g., the question whether a particular mechanism for a phenomenon should be considered) or when it concerns a piece of information for which there is no direct evidence (e.g., the tuning parameter values for a climate model). There exist sophisticated expert judgement elicitation methods that generate this type of evidence in a systematic and reproducible way. The key goal of these techniques is to control for judgement biases that could affect the experts. One of the most prevalent judgement biases is overconfidence. Overconfidence denotes the situation in which an expert is too certain about the truth of her or his statement (Aspinall 2010, p. 294). The crucial point to see is that judgement

[7]Thompson et al. (2016), for example, suggest using the method of structured expert elicitation to incorporate the perspectives of experts into the settling of evidential questions.

elicited in this way then enters into the pool of relevant evidence for a hypothesis. To this pool of relevant evidence, then, one can apply measurement robustness analysis and assess the degree of concordance between the different evidence pieces. Hence, agreement should play an entirely different role than was suggested by the uncertainty framework. It is the agreement of pieces of evidence and not the agreement of experts that is a necessary component for measurement robustness claims.

## 2.6 Ways to improve the uncertainty framework

In Section 2.2, I provided some background information about the structure and aim of the IPCC. In Section 2.3, I offered an interpretation of the uncertainty framework that drew from supporting documents and the practice of the authors. The discussion made clear that the framework contains multiple ambiguities as well as inconsistencies. In Section 2.4, I showed that even a charitable interpretation of the framework faces three substantial conceptual problems. In Section 2.5, I argued for a new way of linking expert judgement and measurement robustness analysis. In my view, the discussion in Sections 2.4 and 2.5 gives concrete leads to how the uncertainty framework should be developed.

Let me now present three steps towards an improvement of the uncertainty framework. First, the bifurcation between evidence and agreement in the confidence metric needs to be removed. As my discussion has illustrated, agreement should not be viewed as an independent dimension from evidence. Rather, our confidence in a finding should be solely determined by the available evidence; the better the available evidence for a finding, the higher our confidence should be in this finding, and vice versa. Making this statement precise is the key task. Importantly, as pointed out before, this is not to say that expert judgement should be erased from the IPCC's uncertainty framework.

Second, assessment criteria for the available evidence need to be identi-

fied and spelled out in sufficient detail. The following questions should be answered: What criteria are relevant for assessing the ensemble of available evidence? What criteria are relevant for assessing individual pieces of evidence? Can one define the assessment criteria in a formally precise way?

Third, once the assessment criteria for evidence have been identified the task of aggregating these criteria into overall confidence judgements can be addressed. Here, the rich literature of social choice theory and multi-criteria decision analysis suggest themselves as sources for technical tools. It remains to be explored to what degree general, non-case specific, aggregation rules can be developed for the context of climate scientific findings.

## 2.7 Concluding remarks

In this chapter, I engaged with the uncertainty framework of the IPCC to address my fourth research question: How should one conceptualise the relation between expert knowledge and the evidence basis when one applies the framework of measurement robustness? At the heart of this uncertainty framework is the notion of robust evidence that is best understood as measurement robustness. The discussion has revealed that the framework exhibits substantial conceptual problems – problems that results from the fact that the relationship between robust evidence and expert agreement is convoluted. I have suggested three ways of improving the uncertainty framework. These suggestions are built around the idea that one needs to ground confidence judgements in evidence and, importantly, that one needs to consider carefully how to incorporate expert judgement into measurement robustness considerations. My answer to the fourth research question is the following: There are two distinct roles for expert judgement with respect to measurement robustness. First, and most importantly, expert judgement can play a crucial role in assessing conditions that need to be in place such that measurement robustness delivers its epistemic goods, and, second, under certain conditions, expert judgement can be a source of evidence.

Note that I have solely taken into account uncertainties involved in the physical science basis of climate change. Ethical uncertainty, which is ad-

dressed in working group III, has been deliberately set aside. In my view, the uncertainty framework should first be improved to address adequately the uncertainties involved in the physical science basis of climate change before it can be generalized to different types of uncertainties.

In the next chapter, I turn to policy-making in the medical domain. This enables me to address issues surrounding derivational robustness analysis.

# Chapter 3

# Robustness in Medical Policy-Making: Automated Large-Scale Evidence Aggregation

## 3.1 Introduction

In biomedical contexts, policy-makers face a large amount of evidence from various sources such as observational studies, randomized control trials, statistical meta-analyses and biochemical studies (Stegenga 2011, p. 497; Krimsky 2005, p. 129; Weed 2005, p. 1545). These sources often confirm conflicting hypotheses. Stegenga (2013, p. 2391-2392) provides an illustrative example: There are three hypotheses about the transmission mechanism of the influenza virus between humans. Although these three hypotheses are not mutually exclusive, understanding their relative contribution to the spread of influenza is critical for designing more effective anti-influenza policies. The droplet hypothesis states that the virus is spread on large droplets that are expelled when people sneeze or cough; the airborne hypothesis states that influenza is transmitted via small airborne particles; the contact hypothesis states that the virus is spread by direct contact between people. Looking at

the literature on these three hypotheses reveals that evidence from a variety of sources has been gathered: evidence from controlled animal experiments using various designs and kinds of animals, mathematical modelling, clinical experience, and epidemiological patterns of influenza outbreaks.

Recently, increased attention has been paid to this situation of evidential diversity. To start, *evidence-based medicine* has moved to the heart of medical policy-making. Originating in the early 1990s, its main purpose is to replace or supplement the potentially biased judgements of clinicians with rigorous evidence-based recommendations. This movement also advocates taking meta-evidence into account. Meta-evidence is evidence concerning how a given set of evidence bears on a hypothesis of interest. Studies reveal, for example, that there is a robust industry bias, that is, medical studies sponsored by the pharmaceutical industry report a higher rate of positive causal efficacy results; or a publication bias, that is, it is more likely that statistically significant results are published than that statistically non-significant results (Cosgrove et al. 2016). Furthermore, medical policy-makers face growing demands for accountability from stakeholders. Taxpayers, regulation authorities, and patient groups ask for transparent decision-making processes that use available resources in an effective manner.

The importance of evidence for medical decision-making prompts a question: How should these large sets of diverse and potentially conflicting evidence, if at all, be aggregated to facilitate best-informed policy recommendations? Throughout this chapter I refer to aggregation procedures involving evidential input of this sort, that is, high volume, highly diverse and potentially conflicting, as 'large-scale' evidence aggregation.

In a series of recent papers, Hunter and Williams have developed a decision support tool to guide medical policy-makers, that is, doctors as well as well people who decide which treatment should be available in a health system (see Gorgoannis et al. 2009; Hunter and Williams 2010; Hunter and Williams 2012; Hunter and Williams 2013). Their framework is used in this chapter to motivate my discussion since it is ambitious and thus raises a number of questions. First, their approach is large-scale and computational – it involves an explicit algorithm, such that, once initiated, little user input is required.

Throughout this chapter, I refer to their approach as 'highly automated'. Second, they provide a full account of the steps involved in choosing a policy recommendation. That is, their framework involves both an epistemic component (assessing hypotheses concerning the various causal effects of the relevant treatments) and a preference-driven component (weighing the treatments in light of these hypotheses). This chapter focuses on the epistemic component. With respect to the epistemic component, Hunter and Williams's proposal raises questions: Does it make sense to deliver a highly automated large-scale evidence aggregator? How does one asses the optimal extent of automation for any given type of evidence aggregation task?

Before I introduce the leading question and the main claims of this chapter, let me add two clarificatory remarks. To start, the notion of *automation* in the context of evidence aggregation needs to be made more precise. For the subsequent discussion, I regard evidence aggregators as automated to the extent that they explicitly encode procedures or algorithms that are fixed, transparent, and dependent on inputs characterised in a particular way. I assume that these procedures are encoded in such a way that they can be run on a computer at high speed, and also so that values for the key parameters of the aggregation function can be easily changed. Non-automated aspects of aggregation procedures refer to the implicit reasoning that affords flexibility, in the sense that how inputs are interpreted and weighed may be determined 'on the fly'. It is effectively expert judgement that may or may not be capable of explicit articulation. So, the main contrast is between reasoning that is made explicit or public, and reasoning that is 'behind the scenes' or happening in a 'black box'. The more the aggregator relies on 'behind the scenes' reasoning, the less automated it is.

Furthermore, this chapter focusses on evidence aggregation regarding general causal claims about the efficacy of treatment options. I am not looking at specific causal claims, that is, claims about the efficacy of treatment options for a particular patient. Accordingly, this chapter does not engage with recent advances in personalised medicine, such as gene sequencing techniques, which allow the assessment of individual biomarkers to develop personalised treatments (Gross 2016, p. xv-xvi). I focus on general causal claims since

they form the content of the evidence aggregation guidelines provided by the National Institute for Health and Care Excellence (NICE) in the UK and a variety of clinical practice guidelines (see Fuller 2013, p. 433-434).

As I show below, the issue of automated evidence aggregation in medicine is a fruitful background against which I can address my second research question: What is the value of derivational robustness analysis if not the entire relevant model space is covered by the available models? As it turns out, structural reflections about evidence aggregation algorithms are the same as those that can be brought up with respect to a set of models. In particular, I argue that to determine the optimal extent of automation, derivational robustness analysis can be a useful tool. Derivational robustness analysis provides a structured way of establishing a preference relation over different algorithm designs by emphasising the amount of relevant possibility space covered by different designs. This can be formulated in the form of a criterion that helps to decide between different algorithm designs. The basic motivation for this criterion is that it directs attention to the question that matters: Does the aggregator have the appropriate input variables and associated parameters? Due to the analogy between algorithms and models, this analysis shows that derivational robustness analysis provides a structured way of forming preference relations over different model structures, where the preference relation is based on the amount of the relevant possibility space that is covered by a model.

This chapter is structured as follows. First, I sketch the general problem of evidence-based comparison of policy options. The leading question for this section is how far standard logic(s) of inference guide the task of constructing a large-scale evidence aggregator. Standard logic of inference helps to frame the aggregation problem, but it does not lead far in answering the question about the optimal extent of automation (*Section 3.2*). I illustrate this with the help of Hunter and Williams's framework (*Section 3.3*). So further criteria are needed for assessing large-scale evidence aggregators. I go on to introduce the ability to perform an adequate robustness analysis as such a criterion, which enables me to state the main claim of the chapter (*Section 3.4*). Finally, I defend my criterion against objections (*Section 3.5*).

## 3.2 Separating components of evidence-based decisions

The claims that are relevant in practical medical contexts are of the sort 'treatment 1 is medically better than treatment 2 for patients of type X'.[1] While such a claim appears to concern only medical facts, it is a claim about preferences – whether treatment 1 is preferred, on all-things-considered medical grounds, to treatment 2.[2] One should not expect to find evidence that bears directly on the overall claim. Rather, behind such a claim is a decision problem that involves the resolution of various value and epistemic issues. The value issues may involve whether one potential health side effect is worse than another, or, more generally, what the more important aspects of health are. The epistemic issues concern the nature and likelihood of the possible health effects of the treatments under consideration.

Consider a person compiling a clinical practice guideline. Assume there are just two treatments for a medical condition.[3] Assume that in the case of the condition under scrutiny both the impact on pain relief and on blood pressure matter. Representing the situation in terms of a multi-criteria decision problem helps to illuminate the components of the problem (see Table 3.1).

| | Pain relief | Blood pressure |
|---|---|---|
| Treatment 1 | | |
| Treatment 2 | | |

Table 3.1: A simplified decision problem between two treatments in light of the criteria 'pain relief' and 'blood pressure'.

Recall that the aim is to assess the claim: 'treatment 1 is more choice-

[1]Let me emphasise that this is a general causal claim not a claim about a particular patient. The patient class description subsumes different conditions that might be in place, such as an age cohort or gender.

[2]Non-medical considerations (such as monetary cost) may also play a role in advice relating to the comparison of medical treatments, but I leave such considerations aside.

[3]For simplicity, I assume that medical conditions, including diseases, can be characterised straightforwardly. For a comprehensive discussion of the nature of such characterisations see Kincaid and McKitrick (2007).

worthy (i.e., medically better, with respect to pain relief and blood pressure), relative to treatment 2, for patients of type X'. Table 3.1 makes explicit that this involves comparing the treatments with respect to both pain relief and blood pressure, and then ultimately ranking the treatments, depending on their performance in these respects, and the relative importance of the two criteria. So, compiling a clinical practice recommendation in this case (e.g., 'For patient of type X given condition C, treatment 1 is recommended') involves both facts (the causal effects of the treatment options with respect to pain and blood pressure) and values (the relative importance of differences in pain relief and blood pressure).

Appreciating the structure of this decision problem is a first, important, step in analysing Hunter and Williams's decision support tool that I outline in Section 3.3. The remainder of this section is concerned with the question of how much further standard (decision) logic can take us in analysing Hunter and Williams's proposals. Besides identifying the role of facts and values in assessing medical hypotheses, one might hope that standard logic of inference is able to give guidance regarding the optimal extent of automation of evidence aggregation processes. I will argue, however, that, although the appropriate logical structure is necessary for an aggregator to provide credible advice, this does not in itself resolve the more substantial question concerning optimal design. I focus mainly on the epistemic aspect of the issue (see Sections 3.2.1-3.2.2) and mention the value aspect briefly at the end of the section (Section 3.2.3).

### 3.2.1 The evidence aggregation problem

For simplicity, in what follows I focus on the comparison of two treatments. In such a situation, the key empirical questions are whether one treatment is better than the other for each respective medical dimension (e.g., pain relief). Accordingly, one wants to know the sign and ideally also the magnitude of the difference in the relevant effect size(s) for the two treatments.[4]

[4]The pain relief effect size, for instance, might be characterised in terms of the proportion of subjects who report a significant decrease in pain after treatment in the treatment group.

In this chapter, I rely on one benchmark logic of inference: Bayesian epistemology. I opt for this logic of inference not to introduce unnecessary technicalities but because this framework allows identifying the components of the inference involved in evidence aggregation in a straightforward and convenient manner. As I shall detail below, Bayesian epistemology forces one to consider individual pieces of evidence as sources of information, the reliability of these sources, and mechanisms used to aggregate the information that takes these reliability assessments into account. In addition, it turns out that the literature on evidence aggregation can be nicely structured with the help of this Bayesian perspective.

In a Bayesian setting, questions of whether and by how much one treatment is better than another translates into a set of hypotheses: either very fine-grained hypotheses concerning the precise difference in effect size between two treatments, or else a more coarse-grained hypothesis concerning simply whether or not treatment 1 is better than treatment 2 along the dimension in question.

The problem of evidence aggregation (this is how I call henceforth the epistemic aspect mentioned in the previous section) concerns what the total evidence at hand says about any given target partition of hypotheses. For the Bayesian, there is a straightforward answer to this question on an abstract level. What ultimately matters is the posterior probability function, denoted here $P^*$, over the hypothesis partition: this is the assessment of the hypotheses upon learning the total evidence. The prior probability function, denoted here $P$, represents the scientist's assessments prior to learning the evidence in question. The posterior probability distribution is equal to the prior conditional probability given the evidence. This expression can be written in ratio form as follows: For some hypothesis $H$ (e.g., 'aspirin is more effective than paracetamol in reducing headaches for patient class Y') relative to its complement, given total evidence amounting to the conjunction of evidence propositions $E_1, ..., E_n$ (Howson and Urbach 2006, p. 20-21):

$$\frac{P^*(H)}{P^*(\neg H)} = \frac{P(H \mid E_1, ..., E_n)}{P(\neg H \mid E_1, ..., E_n)} = \frac{P(E_1, ..., E_n \mid H)}{P(E_1, ..., E_n \mid \neg H)} \frac{P(H)}{P(\neg H)}$$

One can see from the above expression that the change in the prior to posterior probability ratios is driven by the relative likelihoods for the hypotheses, or the relative probability of the total evidence conditional on the respective hypotheses. In the case that $E_1, ..., E_n$ are conditionally independent with respect to $H$, then the likelihood for the total evidence is just the product of their separate likelihoods (Howson and Urbach 2006, p. 18):

$$P(E_1, ..., E_n \mid H) = P(E_1 \mid H)P(E_2 \mid H)...P(E_n \mid H)$$

In the case of dependencies between the evidence propositions[5], however, the likelihoods cannot be thus decomposed.

While abstract Bayesian logic is useful in many respects, it falls short in important respects if one wants to model a real-world problem of evidence aggregation. In particular, the Bayesian model does not in itself provide guidance for determining the relevant likelihood ratio(s) for any given evidence set as well or how to proceed in cases of dependency relations between pieces of evidence. Hence, I propose the following strategy: Instead of trying to fill in a Bayesian model (determining likelihoods and dependencies between evidence propositions) given a description of the aggregation problem at hand, it is easier to adapt one's description of the aggregation problem at hand to a specific Bayesian schema for evidence aggregation.

One useful guiding schema within Bayesian epistemology is one where the individual pieces of evidence are treated as witness reports (see for the following Section 3 of Bovens and Hartmann 2003). A witness report is a statement about the truth of the hypotheses of interest (whether this is a probability of truth or a binary assessment of truth). The likelihood ratio associated with any such report can be regarded as a measure of the reliability of the witness; that is, the ratio of the probability that the witness reports the hypothesis to be true, given that it is true, to the probability that the witness reports the hypothesis to be true given, that the hypothesis is false.

In some more detail, this model can be stated as follows: Let me assume that the evidence propositions $E_1, ..., E_n$ in the expression stated above

[5]This amounts to claiming that $Pr(E_i \mid H) \neq Pr(E_i \mid E_j, H)$, for some $i, j$.

amount to witness reports. In the simplest case, these witness reports concern a partition of just two hypotheses, $H$ and $\neg H$. The reports themselves, denoted $w_1(H), ..., w_n(H)$, each take a value of 1 or 0, the former being a report that $H$ is true while the latter a report that $\neg H$ is true. For $n$ witnesses that are independent (conditional on the hypotheses in question), the likelihood ratio, which expresses the impact of the evidence on prior beliefs about the hypotheses, is as follows:

$$\prod_{i=1}^{n} \frac{P(w_i(H) \mid H)}{P(w_i(H) \mid \neg H)}$$

Note that each term in this product amounts to the likelihood ratio for the witness report for the pair of hypotheses. This is effectively the reliability (or else its inverse) of the witness with respect to her report in favour (or against) the truth of $H$. So, one sees that the impact of the testimonial evidence of a number of independent witnesses, in the Bayesian schema, is a fairly simple aggregate (namely a product) of their respective report reliabilities (or inverse reliabilities).

The task of applying the witness schema to a real case consists in identifying what the separate pieces of evidence are that may be modelled as giving independent witness reports. Then, the reliability of these reports need to be assessed. So, in conclusion, what one gains from abstract Bayesian logic is a tractable way of framing an evidence aggregation problem: one needs to identify sources of information, assess their reliability, and come up with an aggregation mechanism. As discussion of the existing literature on evidence aggregation below and Hunter and Williams's framework in the next section will make clear, however, this is not enough to answer the question regarding the optimal extent of automated evidence aggregation.

### 3.2.2 The literature on evidence aggregation vis-a-vis the schema

The existing literature on aggregating diverse evidence for causal hypotheses fits well with the witness schema introduced in the previous section. That is

to say that much of the literature can be interpreted as being about the more substantial questions emerging if one tries to apply this schema to concrete cases: How should individual witnesses (and their reports) be delineated to preserve independence and how should the reliability of these reports be assessed? What algorithm should be used for aggregating the reported causal conclusion?

The evidence-based medicine hierarchies are a good case in point for the preoccupation in the literature with determining the reliability of (types of) witnesses and the aggregation mechanisms of their reports (see Clarke et al. 2014 for an overview). For instance, the hierarchy of evidence introduced by NICE places meta-analyses, systematic reviews of RCTs, and RCTs above case-control or cohort studies, non-analytical studies such as case reports, and expert opinion (NICE 2006). Hierarchies of this sort clearly can be interpreted as making claims about the reliability and aggregation of evidence from different types of sources. The relative standing of a (type of) witness, such as RCTs, is justified with reference to their respective reliability assessed based on study type design (e.g., the sample sizes and the controls for bias). RCTs are often claimed to be the gold standard to establish causal relationships between variables and, more precisely, to determine the size of the causal effect of a particular intervention (see the overview discussion in Cartwright and Hardie 2012). The large body of literature which criticises the top-ranking spot of RCTs proceeds by providing arguments for why RCTs might not be reliable in predicting the effectiveness of interventions in a given context. For example, Kincaid (2001, p. 36) points out that the inclusion criteria of RCTs are often so restrictive that it is not clear what RCT results reveal about a general population or whether randomisation techniques provide assurance that in a particular trial (in contrast to the repetition of the same trial design) the control and treatment group are balanced.

The evidence-based medicine hierarchies are furthermore a good case in point for discussing the ambition in the literature to assess the independence of witness reports. A major criticism of the evidence hierarchies is that mechanistic evidence and expert opinion are placed at the bottom. Based on the Russo-Williamson thesis, that is, the idea that establishing a causal relation-

ship requires observing probabilistic association and stating a mechanism for the association (Russo and Williamson 2007), the claim is that mechanistic evidence for causation is rather complementary to statistical evidence. Instead of undermining the witness schema, this debate sounds a note of caution regarding what counts as independent witness-style evidence. In particular, one might argue that evidence of mechanisms and evidence of probabilistic association cannot be treated as separate witnesses.[6]

One might worry that the witness model does not account for statistical meta-analysis. Statistical meta-analysis involves four steps: a) selecting the primary studies, b) determining the appropriate outcome measure for each study (such as effect size), c) weighing each study (usually according to its size and quality, for example by using the inverse of the variance of the effect estimate), and d) calculating the weighted average of the effect sizes across the studies (Stegenga 2011, p. 497-498). However, there are two ways of reconciling statistical meta-analysis with the witness schema. The first is to see meta-analysis as an instance of the witness schema. In this case, statistical meta-analysis chooses a particular algorithm for aggregating witness reports and is limited to a certain kind of evidence (e.g., fine-grained effect size results from RCTs). The second is to treat an individual meta-analysis as a single study (or its results as a witness report). This might be viewed as a particularly reliable study since it combines data from multiple studies (but see Stegenga 2011 for doubts about the reliability of meta-analyses). In what follows, I take the latter route, treating meta-analyses as a single study, because I am interested in attempts at large-scale evidence aggregation, that is, involving more than just evidence from RCTs.[7]

One might also wonder how more complex methods of causal inference

[6]To keep the discussion focussed, I do not engage here with the question of whether the Russo-Williamson thesis holds. As Kincaid (2011) points out, to assess this claim one needs to distinguish carefully between different meanings of the term mechanism and whether one assesses hypotheses about the existence of a causal relation or the effect size of a cause.

[7]Let me point out that evidence guidelines, which review existing evidence for a condition, usually treat meta-analyses as the final aggregation products (see Fuller 2013, p. 436). This can be explained by the fact that meta-analyses of RCTs are ranked the highest in evidence-based medicine hierarchies, and, hence, are usually regarded as trumping other evidence types.

can be reconciled with the witness schema. Bayesian nets methods involve inferring a causal graph from a joint probability distribution over a number of variables (see for an introduction and discussion Ben-Gal 2007). Here again, one might treat any such study as a single witness report with respect to the particular cause and effect. Given the ability of these methods to identify confounding and/or mediating variables, one might judge this source of information as highly reliable. While the evidence aggregation problem considered in this chapter is large-scale when it comes to the diversity of evidence, it is in another sense simple in that one restricts attention to a single cause and effect relationship, as opposed to a more detailed and complex causal network. Accordingly, I am going to set these Bayesian net methods aside.

In general, the above discussion supports the point that the witness schema is a very flexible model for evidence aggregation. Clearly a lot depends on how the evidence is divided into independent witnesses. That is, a lot of the difficult questions in the context of evidence aggregation are shifted to the 'pre-processing' of evidence and, hence, away from the final aggregation task. So what one gains from abstract Bayesian logic is that one should assess any proposal for aggregating evidence according to the way the evidence is delineated and scored for reliability, in addition to how it is processed in the aggregation function. This is an additional insight over the one gained by seeing evidence as a multi-criteria decision problem involving an epistemic and value-based component. In other words, the bulk of the aggregation task is not settled by an abstract inference logic. How much automation is a good idea is a question that remains unanswered.

### 3.2.3 Return to the multi-criteria decision problem

As noted earlier, in order to finally make a choice of treatment, one must not only tackle the evidence aggregation problem for the empirical hypotheses, but also the problem of combining the various empirical and evaluative claims that are relevant for the final choice of options. This is a contested issue and falls under the literature of multi-criteria decision analysis (for an overview

see Keeney and Raiffa 1993). There are a host of suggested methods for making a choice in the face of options that are ranked differently according to different criteria which themselves have differing importance for the decision-maker. An important specification of these methods is whether the ranking of options for each dimension can be represented in cardinal terms or only in ordinal terms; that is, can one specify how much better/worse treatment 1 is on pain relief compared to treatment 2 for patient type X or can one only say which is better/worse? When cardinal information is plausibly available and the criteria are comparable in cardinal terms, the decision analysis can follow standard Bayesian principles in the form of expected utility theory. When only ordinal information is available, the controversies regarding how to make an overall choice are more extensive (see Chapter 6 for discussion of the Arrovian problem for aggregating ordinal information).

## 3.3 Hunter and Williams's highly automated evidence aggregator

Hunter and Williams claim to offer "a new framework for representing and synthesizing knowledge from clinical trials involving multiple outcome indicators" (Hunter and Williams 2012, p. 1). In line with what I said in the introduction, Hunter and Williams's proposal should be understood as a tool to arrive at general causal claims. Here, I summarise what I take to be the main tenets of their account. The key upshot of the discussion is that their procedure illustrates the claim made in the previous section: to answer the question about the optimal extent of automation of an evidence aggregator, further criteria are needed over and above the ones provided by standard logic of inference.

The goal of Hunter and Williams's approach is to come up with an ordinal ranking of two treatment options (Hunter and Williams 2013, p. 16).[8] For example: Is the treatment of contraceptive pill or no treatment better,

[8]I discuss solely the case involving two treatment options. Hunter and Williams's proposal is able to deal with multiple pairwise comparisons of treatments.

when one looks at pregnancy, ovarian cancer, and breast cancer as outcome indicators for patient class $Y$? (Hunter and Williams 2013, p. 16)[9]

In line with my discussion of the decision problem underlying evidence aggregation, their approach contains two aggregation elements:

1. The aggregation of evidence for each outcome indicator, which involves a) delineating evidence, b) assessing reliability of evidence, and c) an aggregation rule.

2. The overall multi-criteria problem, taking into account all relevant outcome indicators.

Given the focus of my discussion, I explain the first of these two aggregation steps in more detail. I briefly sketch their approach to the second step for completeness. During my exposition of their proposal, I will point out its automated nature and establish that it is a partly automated aggregation procedure.

To aggregate the available evidence regarding the various outcome indicators, Hunter and Williams introduce an evidence table. This table delineates evidence in terms of different studies. For each study, a number of key input variables are filled in or evaluated (Hunter and Williams 2012, p. 5-7):

- the pair of treatments (e.g., contraceptive pill, no treatment)

- the outcome indicator denoting the dimension along which the treatments are compared (e.g., breast cancer)

- the value of the outcome indicator given a particular measure. If one adopts the relative risk measure, then the ratio between the portion of people displaying the outcome, given treatment 1 (e.g., breast cancer given contraceptive pill), and the portion of people displaying the outcome, given treatment 2 (e.g., breast cancer given no treatment), is calculated

[9]Ultimately, Hunter and Williams's proposal allows treatment recommendations for individual patients given the individual patient falls into the relevant patient class (Hunter and Williams 2010, p. 119).

- the net outcome indicating whether $T_1$ is superior, inferior, or equivalent to $T_2$. The net outcome is determined by the measured value of the outcome indicator and whether the outcome indicator is desirable for a patient class. For example, given a value of 1.04 (assuming the above specified measure of relative risk and the outcome indicator breast cancer), the contraceptive pill is inferior to no treatment (Hunter and Williams 2012, p. 13)

- whether the measured value of the outcome indicator is statistically significant

- the evidence type (e.g., RCT study, cohort study, meta-analysis, or network analysis)[10]

These input variables inform what the evidence says about the treatments and how reliable it is. The information in the evidence table is the total available input for the automated aggregation procedure. Their algorithm can "be run" based on this information, given some additional specifications that I introduce below.

The different studies in the evidence table are treated as independent witnesses. For each witness, its reliability is determined via so-called meta-arguments. Hunter and Williams consider the following meta-arguments (for identifying unreliable evidence): 'the evidence contains flawed RCTs', 'the evidence contains results that are not statistically significant', 'the evidence is from trials that are for a very narrow patient class', 'the evidence has

[10]This is merely one candidate list for characterising evidence. Hunter and Williams note that further information should be captured by the evidence table if practicable; for example, the sample size of a trial, the geographical location for each trial, the drop-out rate, the method of randomization, and whether a trial used a narrow patient class (Hunter and Williams 2012, p. 7). However, note that they only aggregate clinical data. Non-clinical data, such as genomic information about the patient classes under consideration are not considered. The approach via the evidence table also presupposes that the problem of data integration, that is, how data from different sources can be made compatible (Leonelli 2013), is solved.

outcomes that are not consistent'.[11] These meta-arguments return binary 'in or out' results (meaning that a study either feeds into the aggregation function or it does not) for computational ease and theoretical simplicity (Hunter and Williams 2012, p. 4, 9, 20).

With these elements in place, one can discuss Hunter and Williams's aggregation rule for any given outcome indicator. It is based on the notion of an inductive argument, where this is a pair $(X, \epsilon)_i$, with $X$ being a subset of evidence and $\epsilon$ the claim that either $T_1$ is superior to $T_2$, $T_1$ is inferior to $T_2$, or $T_1$ is equivalent to $T_2$ with respect to outcome indicator $i$ (Hunter and Williams 2012, p. 3, 4, 11).[12] The set of inductive arguments is constructed by the algorithm in the following way. To start, the available evidence (concerning $T_1$ and $T_2$) is divided into three subsets SUPERIOR, INFERIOR, and EQUITABLE. The subset SUPERIOR contains all rows of the evidence table (i.e., evidence pieces) for which $T_1$ was shown to be superior to $T_2$. The subsets INFERIOR and EQUITABLE are defined via the inferiority and equivalence relations, respectively (Hunter and Williams 2012, p. 10-11). The inductive arguments are determined by permissible inference rules from the set of evidence. Hunter and Williams propose the following three simple inference rules that apply to cases where evidence is not conflicting (Hunter and Williams 2012, p. 11):

- If $X \subseteq$ SUPERIOR, then $T_1 > T_2$
- If $X \subseteq$ INFERIOR, then $T_1 < T_2$
- If $X \subseteq$ EQUITABLE, then $T_1 \sim T_2$

For example, given that two RCTs in the evidence table $(E_1, E_2)$ state that the contraceptive pill $(T_1)$ is inferior to no treatment $(T_2)$ for the outcome

[11] Hunter and Williams do not claim that this list of meta-arguments is definitive. Different meta-arguments could be used in different application contexts or if different clinicians use the aggregator for the same problem because the notion of 'reliable evidence' is debatable (Hunter and Williams 2012, p. 25). Furthermore, Hunter and Williams explicitly say that this flexibility of their aggregator allows for "a form of sensitivity analysis". I take up this idea in Section 3.4.

[12] I have augmented Hunter and Williams' notation for an inductive argument to indicate that inductive arguments are indexed to outcome indicators.

indicator breast cancer ($k$), the inductive arguments $(E_1, T_1 < T_2)_k$, $(E_2, T_1 < T_2)_k$, and $((E_1, E_2), T_1 < T_2)_k$ can be generated. In a nutshell, their procedure can be viewed as taking the last of one of these inductive arguments as being the relevant one; that is, the inductive argument with the broadest set of evidence for the respective superiority, inferiority, or equitability claim for an outcome indicator.

With the notion of an inductive argument in place, one can specify the aggregation rule for a single outcome indicator. Consider the example of two studies with contradictory claims about the effects of contraceptive pills. Let me assume that two inductive arguments can be generated for the outcome indicator ovarian cancer ($o$): $(E_1, T_1 < T_2)_o$ and $(E_2, T_1 > T_2)_o$, where $T_1$ is the contraceptive pill and $T_2$ is no treatment. Hunter and Williams propose two options to aggregate this information.[13] First, they suggest performing a statistical meta-analysis which aggregates the effect sizes of the two studies. Once the effect size are combined one then can infer the binary superiority relation. This amounts to generating a new inductive argument (while deleting the two other ones) with the evidence basis $E_3$ being the meta-analysis: $(E_3, T_1 < T_2)_o$. This corresponds to laundering the evidence such that it no longer conflicts. Second, the reliability criterion could be contradicting one (or both) of the inductive arguments and thereby resolve the conflict between the two arguments. It seems that Hunter and Williams assume that all (or certainly the largest part) of the studies will give the same ordering for the treatments for any particular indicator. Note that they do not propose an alternative aggregation rule which is salient: a majority (or supermajority) rule could be used in assessing the relevant majority proportions between numbers of studies indicating that $T_1$ is superior (inferior) to $T_2$. This aggregation algorithm shows that some expert judgement is required in the case of conflicting evidence. Hence, their procedure is best described as partly automated.

So far, I have described the first aggregation step. For the sake of completion, I turn to their treatment of the overall multi-criteria problem: What is the best treatment option taking into account all relevant outcome indicators? Hunter and Williams suggest a pairwise comparison of the inductive

[13]This was clarified in personal communication with Hunter and Williams.

arguments involving different outcome indicators: for example, the comparison of two inductive arguments one stating that $T_1$ is superior to $T_2$ with respect to ovarian cancer, the other stating that $T_2$ is superior to $T_1$ with respect to pregnancy (Hunter and Williams 2012, p. 15).

To this end, Hunter and Williams introduce a preference relation that is defined over inductive arguments. The preference relation is derived from a notion of benefits of treatments. Benefits of a treatment are defined by the outcome indicators and their values: considerations regarding the relative standing of the outcome indicators (e.g., pregnancy or breast cancer) and the respective relative risks of developing the outcome indicators (e.g., small or great relative risk reduction) (Hunter and Williams 2012, p. 15). To illustrate the notion of benefit of treatment and how it can be used to construct a preference relation over inductive arguments, consider the following hypothetical example. Two evidence pieces for the same treatment pairs (e.g., $T_1$: contraceptive pill, $T_2$: no treatment) are available. Study 1 ($E_1$) reports for the outcome indicator 'breast cancer' ($k$) that $T_1$ is worse than $T_2$; study 2 ($E_2$) reports for the outcome indicator 'pregnancy' ($l$) that $T_1$ is better than $T_2$. Based on these studies, the following two inductive arguments can be constructed: $(E_1, T_1 < T_2)_k$ and $(E_2, T_1 > T_2)_l$. Note that the two outcome indicators pull in different directions. In this case, Hunter and Williams suggest that we should prefer the benefit of the substantially reduced risk of becoming pregnant to the disadvantage of having a slightly increased risk of developing breast cancer (Hunter and Williams 2012, p. 15). Hence, the preference relation between the two inductive arguments looks like this:[14]

$$(E_2, T_1 > T_2)_l \succ (E_1, T_1 < T_2)_k$$[15]

[14]They characterize the preference relation by a set of properties (Hunter and Williams 2012, p. 17-18). These properties, for example transitivity, can be used to extend the preference relation to pairwise comparisons of treatment options for which benefits have not been explicitly determined.

[15]One might worry that this suggestion runs into the problem of overlooking confounding factors. For, is it not the case that one can only make statements about the benefits of the two treatments in terms of 'pregnancy' and 'breast cancer' if the two treatments have been evaluated in the same study? Hunter and Williams account for this problem via meta-arguments that ensure that the patient classes are similar enough such that these comparisons of different studies are possible.

Hunter and Williams make suggestions about how the construction of the preference relation could be automated. One way to do it, is to specify rules for its construction; for example, the rule that for any outcome indicator concerning survival and any minor side effect, the inductive argument involving survival is preferred to the inductive argument involving the minor side effect, if the treatment increases the chances of survival and the increase in the side effect is of the same order of magnitude as the increase in survival (Hunter and Williams 2012, p. 17). If such a set of rules is implemented, the algorithm can execute them and construct the preference relations over the inductive arguments.

Once such a preference relation is defined over all inductive arguments, the multi-criteria decision problem can be resolved by determining what Hunter and Williams call the winning inductive argument. The winning inductive argument is determined by the algorithm as the inductive argument that does not appear in any preference relation as being preferred over another inductive argument. If the two inductive arguments in my hypothetical example involving breast cancer and pregnancy were the only two inductive arguments under consideration, $(E_2, T_1 > T_2)_l$ would be the winning argument. Once the winning argument is identified, the all-things considered preference relation between the treatments can be read off straightforwardly. In my hypothetical example, one could read off the ordinal ranking $T_1 > T_2$.[16]

Let me take stock here. Hunter and Williams put forward a decision support tool that has the components outlined in Section 3.2. The evidence aggregation component can be seen to accord with the Bayesian witness schema. Thus, the right building blocks are present in Hunter and Williams's aggregator. However, a lot of design choices in their aggregator are not settled by the abstract inputs of a Bayesian decision logic. In fact, there are many design choices in Hunter and Williams's aggregator that one might challenge. For example, Hunter and Williams's evidence aggregator can strike one as rather crude; evidence is delineated simply in terms of separate studies; there is only

[16]Note that the set of winning arguments does not need to be a singleton. This is not a problem if all the winning arguments express the same ordinal ranking of the treatment options. If this is not the case, Hunter and Williams's procedure yields an indifference relation between the two treatment options (Hunter and Williams 2012, p. 20).

a small set of meta-arguments that are used to form the reliability judgements; and the binary reliability judgements do not allow ruling in favour of one evidence piece over another in degrees. But maybe this crude design of the aggregator does strike the right balance between various desiderata, and, hence, corresponds to the optimal amount of automation. How should one even begin to make such an assessment? I will turn to this issue in the next section.

## 3.4 A new assessment criterion for determining the optimal degree of automation: Capacity for robustness analysis

Similar to other approaches to evidence aggregation, Hunter and Williams's proposal can be seen to accord with a witness model, but this in itself does not go far in terms of settling the quality of the inferences. I now consider in more detail whether their procedure is fit for purpose. I address this issue not solely for Hunter and Williams's decision support tool but with respect to automated evidence aggregation procedures more generally: What is the appropriate extent of automation for an evidence aggregator to help facilitate policy recommendations in a medical context?

The large question that is left open by the witness schema is how to delineate the independent witnesses, and assess the nature and reliability of their findings. Reliability, for instance, is a matter of both the quality and relevance of the experiment or witness, given the hypotheses at hand. One must determine what features of individual pieces of evidence should inform this complex reliability assessment and how exactly reliability should be determined on the basis of these features. Assessing the reliability of the witnesses is one of the key tasks that is automated in automated evidence aggregators. Recall that Hunter and Williams automate the reliability assessment via meta-arguments.

In Section 3.4.1, I consider this question first from an ideal perspective, free from any constraints on computational resources. In effect, I consider better and worse treatments of (or inferences based on) a fixed amount of

evidence (the more the better). Here I introduce the key idea of ability to perform an adequate robustness analysis. In Section 3.4.2, I consider the more realistic scenario where there may be constraints on computational resources. I show that this is a further problem of a trade-off between nuanced analysis of evidence and volume of evidence. I argue that it may not be the case that aggregators that can handle more evidence are better. I will argue that there is no requirement to aggregate the total evidence available if this would in fact reduce overall accuracy.

### 3.4.1 Assessing aggregators in an ideal setting: no computational constraints

In the absence of any computational constraints, the goodness of an automated evidence aggregator is all about ideal performance – the idea is to make the best or wisest inferences possible for a class of cases, given all the available evidence. Thus, one wants the reasoning process to be as nuanced as possible, where the more of this nuance that can be captured by an explicit algorithm, and so automated, the better. A higher degree of automation is desirable, all else being equal, because it allows transparency, removes computational error, enhances speed, and facilitates analysis of the sensitivity of results to choices of parameter values, that is, a robustness analysis. In the ideal setting then, any part of the reasoning process that can be made explicit in advance of seeing the particular evidence at hand, should indeed be made explicit. The only reason to leave some aspects of the reasoning process as a black box is if this is advantageous in terms of the accuracy of the inference. In this case, it is better to leave it to experts to interpret and weigh the particular evidence when it arises.

It is one thing to state the goal of automating all reasoning that can be made explicit without sacrifice in analysis and that improves the accuracy of the aggregation, but it is another thing to make these extremely difficult judgements. To give detailed advice on how to make such judgement would be to ask too much from this chapter. For one thing, much of the detail will depend on the type of policy task at hand and the kind of evidence available.

Instead, what I am looking for are strategies that a practitioner may employ to approach or frame the question in a way that may assist in arriving at an answer. To put it differently: What general criterion provides the best avenue for assessing the optimal degree of automation?

The criterion I propose for assessing the degree of automation of an evidence aggregator is: Does the automated aggregator permit one to conduct a robustness analysis that would yield a thorough and compelling survey of the possibility space? Note that robustness analysis has already been mentioned above as a useful byproduct of automation. I cover this point briefly in the next subsection. The novel proposal, however, is that this consequence of automation can serve as a key point in the algorithm design. This is a matter of deciding the explicit algorithm structure ex ante with an eye to whether the subsequent robustness analysis will serve as a reasonable survey of the possibility space for the type of aggregation problem at hand. In short, focussing on the ability to conduct an adequate robustness analysis serves to direct one's priorities to what really matters – away from the precise 'dial settings' of an aggregator, so to speak, and towards whether one has the right dials to begin with.

Before proceeding, let me first clarify certain terms concerning the structure of an automated evidence aggregator. To start, there are *input variables* describing the evidence. Recall, for instance, Hunter and Williams's evidence table: the columns are the input variables accounting for relevant features of the evidence, and each row – an individual piece of evidence – is effectively a vector of values for these input variables. Furthermore, there are the *parameters* of the aggregation function that dictate how the values of the input variables bear on the assessment of each piece of evidence and ultimately on the overall aggregation or final inference concerning the hypotheses in question. For instance, the aggregation function might include a parameter 'threshold sample size', which is used to measure the quality of a piece of evidence with respect to sample size. Finally, these parameters are associated with a range of *possible values*. Thanks to programming design, the parameters can be set to any value within this range/set, depending on initial user input.

### Robustness analysis in its traditional role – as a useful byproduct of automation

As we saw in the introduction to this thesis, robustness analysis involves determining the stability of a result given changes in assumptions. I distinguished two types of robustness analysis: Derivational robustness analysis looks at the stability of model derivations given changes in the model assumptions (Kuorikoski et al. 2010, p. 542). Measurement robustness analysis looks at the stability of empirical results given changes in empirical modes of determination (such as different types of experiments) (Wimsatt 1981, p. 128).

Here, I focus on derivational robustness analysis for two reasons. First, I do not want to assume that (or assess whether) the various pieces of evidence indicate the same ordering of treatments, as would be the focus of measurement robustness analysis. In fact, as Stegenga (2009) convincingly argues, in the biomedical sciences one is usually facing discordant evidence. As I point out in the next section, it is not the robustness of the results but the ability to perform a robustness analysis that is the focus of my new criterion. Second, derivational robustness analysis is a form of error analysis, that is, a way of exploring the sensitivity of results to choices of parameter values. This error analysis can be read in a heuristic way. In particular, derivational robustness analysis allows a transparent and traceable way of dealing with unavoidable idiosyncratic choices in the construction of an evidence aggregator. These parameter value discrepancies might occur through uncertainty about the correct values for an epistemic agent or through reasonable disagreement between epistemic agents. Derivational robustness analysis can play this role since it allows, as pointed out in the introduction, determining the relative importance of various assumptions in relation to the result (Kuorikoski et al. 2010, p. 543).

Given my assumption about how an automated evidence aggregator is implemented, these aggregators are well set up for this kind of error analysis. This point has been brought up by Hunter and Williams in relation to the meta-arguments in their aggregator. They note that their procedure allows "a form of sensitivity analysis" by including different meta-arguments

(Hunter and Williams 2012, p. 25). By including different meta-arguments, the reliability of the evidence is assessed differently.

In other contexts too, sensitivity analysis is recommended as a way to keep track and explore the implications of model choices that are subject to uncertainty and/or reasonable disagreement. For instance, Stegenga (2011, p. 498) points out that there are many choices of this kind in statistical meta-analysis:

> Meta-analysis fails to constrain intersubjective assessments of hypotheses because numerous decisions must be made when performing a meta-analysis which allow wide latitude for subjective idiosyncrasies to influence the results of a meta-analysis.

Let me now turn to the new role of robustness analysis in the context of my criterion for assessing the optimal extent of automation for evidence aggregation procedures.

### Robustness analysis in its new role – as a central design criterion

For all the good of derivational robustness analysis, one might regard it a secondary issue when it comes to assessing the degree of automation of an evidence aggregator. Surely the primary issue is whether the aggregator facilitates roughly the best inferences possible given the available evidence; error analysis is a matter of extra detail. As suggested in the previous section, robustness is indeed typically considered an ex post analysis or a way to check what confidence one should have in a model result. Here I want to defend, however, a more central role for robustness analysis in the construction and assessment of an evidence aggregator. In short, the prospect of what robustness analysis can be performed focuses one's attention on what really matters, that is, the functional form and possible inputs to the evidence aggregator, rather than the precise parameter values featuring in the aggregator. Put differently, viewing the construction of an automated evidence aggregator through the lense of robustness analysis helps one to assess what parts of the inference process can be made explicit and transparent.

There are two reasons why focussing ex ante on the capacity for robustness analysis is helpful in making these judgements about algorithm design: To start, it allows one to recognise that certain types of uncertainty or error regarding precise parameter values do not compromise automation, since the impact of these uncertainties can be explored via the robustness analysis ex post. Furthermore, it allows one to recognise that other types of uncertainty or error do in fact compromise automation, and, hence, call for a lesser extent of automation. These are cases where there is not only low confidence in the 'best guess' estimates for parameter values, but where there is low confidence in the entire possibility space that would be afforded by robustness analysis accomplishable with the algorithm design under scrutiny. In this case, it is not clear whether the remedy is more or less automation, but robustness analysis can guide the deliberation process.

The basic idea of the latter point can be visualised with the help of a diagram (see Figure 3.1). Assume that you face a particular evidence aggregation task (e.g., the assessment of the causal efficacy of a treatment with respect to pain relief and blood pressure). You ask yourself what optimal extent of automation of an aggregator is for this type of aggregation task. The criterion introduced above asks you to make this decision in the light of robustness considerations. In particular, it asks you to choose an aggregator design that allows better coverage of the relevant possibility space. In Figure 3.1 this would be aggregator design $B$. The relevant possibility space (illustrated with the help of a box) contains inferences (and their outcomes) that are deemed relevant for the evidence aggregation problem at hand. Although the space is here visualised in two dimensions, the space is in fact multi-dimensional, being created by variations in input variables, forms of the aggregation function defined over these input variables, as well as parameter values associated with the input variables. For example, one might be unsure whether for a particular aggregation task the sample size of a trial matters or not and how it should matter; multiple stances can be taken in relation to this question and, hence, a space of relevant possibilities opens up. Importantly, the qualification 'relevant' should not be understood as denoting the *theoretically* relevant space, that is, all epistemically defensible variations

of an aggregation procedure for a given task at hand. Rather, it is a *practically* relevant space, that is, the space of possibilities that are entertained by the epistemic agent faced with the aggregation task.

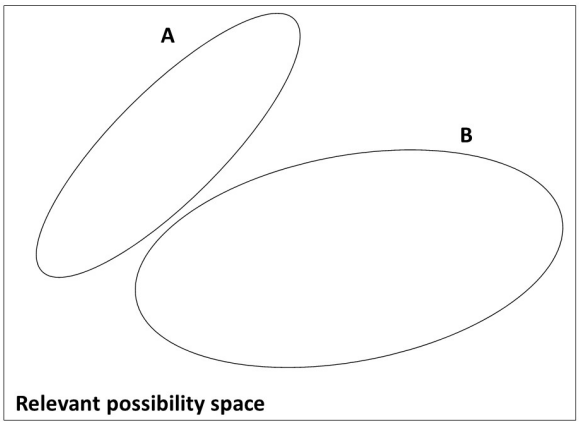


Figure 3.1: Two different aggregator designs $A$ and $B$ that entail different coverages of the relevant possibility space (my diagram).

Let me clarify the criterion with an example. Imagine the incremental development of an evidence aggregator. The starting structure might be a very basic one, where the pieces of evidence are described in terms of two input variables, say, 'type of study', with possible values 'randomised controlled trial (RCT)' and 'observational study', and also 'statistical significance', where possible values are simply 'yes' and 'no'. The logic of the reliability assessments might proceed along the following lines: Only studies that are statistically significant have positive reliability (such that they are included in the aggregation), and amongst those, the RCTs are given more weight according to a parameter $\beta$, specifically, RCTs are given $\beta$-times the reliability weighting of observational studies. Note that Hunter and Williams introduce a crude reliability judgement of this sort by considering meta-arguments that include/exclude evidence pieces based on whether results are statistically significant or not. Now, one might reflect on the robustness analysis afforded by this aggregator design. The possibility space, which can be covered by a robustness analysis, will include inferences based on a range of values for $\beta$. But this might be regarded too limited a set of possibilities.

For the above example aggregator, the possibility space afforded by robustness analysis might be deemed more adequate if the algorithm for making reliability judgements were more detailed. A relatively straightforward innovation is to convert judgements that are currently implicit but need not be explicit aspects of the algorithm. For instance, with regard to the example, the judgements of statistical significance could be spelled out more explicitly. One way of doing this is substituting the p-value of the study as the input variable, and then deriving whether the study is statistically significant according to the parameter $\alpha$, such that if the p-value is less than $\alpha$, the study is deemed statistically significant. The corresponding robustness analysis would then produce a possibility space that includes a range of values for $\alpha$, which would presumably be more thorough.[17]

The more difficult judgements concern aspects of the reasoning process where it is not clear whether more or less detail in the explicit algorithm would be better. Adding detail to the explicit algorithm is a good thing provided this is tracking a genuine nuance of reasoning. However, there is an alternative scenario in which extra detail in the explicit algorithm systematically distorts the reasoning process. This can happen, for example, by making the algorithm design more rigid in a way that is not rectified by robustness analysis. Returning to the example, a key reason why the initial robustness analysis might be deemed inadequate is that the reliability weightings depend purely on 'study type', and it might be thought that this is not the most pertinent grouping as far as quality of evidence is concerned. One possibility is to add further dimensions to this grouping: perhaps 'sample size' and a measure of the 'relevance of experimental subject', that is, the closeness of the experimental group to the patient class at hand, could also be included as an input variable, and treated in the reliability function with reference to appropriate parameters. In this case, the robustness analysis would effec-

[17]Let me be clear at this point that I do not want to endorse statistical significance as important for determining whether a study result ought to be included in evidence aggregation. There are reasons to worry about this interpretation of statistical results. The claim is simply that if this property were to play such a role, better to make the reasoning as explicit as possible, and enable the exploration of changes in the chosen level of significance.

tively survey the possibilities associated with changing the relative weights of these more fine-grained study groupings, which would potentially be a more adequate representation of the real space of possibilities.[18]

At this point, I can take stock and connect the discussion to the leading question of this chapter: What is the value of derivational robustness analysis if not the entire relevant model space is covered by the available models? As my discussion in this section was supposed to show, the structural reflections about evidence aggregation algorithms are the same as the one that can be asked about a set of models. There too exists a situation in which an epistemic agent must judge how much of the relevant possibility space is covered. The transferable upshot of the discussion in this section is that, in cases in which not the entire relevant possibility space is covered, derivational robustness analysis nevertheless can play a fruitful role: it provides a structured way of forming preference relations over different model structures, where the preference relation is based on the amount of the relevant possibility space that is covered by a model.

### 3.4.2 Computational constraints: When is greater volume of evidence better?

I turn now to the scenario where there are constraints on computational and other resources. In practice this is always the case. Hunter and Williams also have resource constraints in mind when they motivate the usage of their aggregator (Hunter and Williams 2012, p. 2). To start, the on-going maintenance of the database of evidential inputs demands a lot of person hours for identifying and recording the relevant features of each experiment or piece of evidence. When it comes to running the aggregator, once the database is in hand, the algorithms for selecting, assessing the reliability and aggregating relevant pieces of evidence to arrive at a conclusion about the hypotheses in

[18]For example, one can define for the input variable 'sample size' the parameter $\gamma$ such that the reliability assessment of a study correlates not linearly to the sample size but increases in a step-function fashion. For the input variable 'relevance of experimental subject', a parameter $\delta$ could be introduced in relation to the similarity measure between the experimental and patient group.

question all require processing time. In short, the aggregator may be subject to various limitations on resources. The assessment of evidence aggregators in the presence of such constraints calls for a trade-off between inferential accuracy and efficiency.

The way forward is to carefully think about how one should spend one's budget to best achieve inferential accuracy. It is not the case that all available evidence must be taken into account. The tenet that inference must be based on all available evidence[19] refers to an ideal setting but not to a practical setting where, due to resource constraints, this evidence cannot all be assessed in full detail. That would indeed be an odd requirement on an evidence aggregator – that just because some apparently relevant evidence has been tabled, it must influence the inference at hand, even if in a necessarily crude fashion.

My claim in relation to the question of the optimal extent of automation given resource constraints is the following: The assessment of an evidence aggregator under circumstances of resource constraints is not so different from the assessment of an aggregator that is free from resource constraints. In both cases it is performance, that is, quality of inference regarding the hypotheses at hand, that matters. Once again this is best assessed by focussing on the capacity for robustness analysis. In the context of resource constraints, the 'principle of total evidence' may be better honoured by processing a subset of evidence in more detail rather than a greater amount of evidence in lesser detail.

This further balancing act can be informed by the capacity of an evidence aggregator for robustness analysis. Let me discuss an extreme scenario that may shed light on the more difficult non-extreme cases. The worst case, is an evidence aggregator for which one is not confident that any of the input evidence contributes to higher quality inference about the hypotheses at hand, regardless of the precise values of key parameters. Here, it is not even the case that, conditional on using the aggregator at hand, inferences based on more evidence are better. In the more ordinary and difficult cases, by contrast, all the aggregators under assessment will be ones for which more evidence

[19]This is the *Principle of Total Evidence* (Carnap 1947).

(requiring more resources) permits better inferences using that aggregator. Here, the possibility space associated with each aggregator, owing to robustness analysis, is at least deemed adequate. The further question in this case is which aggregator allows for the best quality inferences given the resource budget at hand, where some aggregators process less evidence with greater nuance while others process more evidence with lesser nuance. Thus, what using robustness analysis as a lens allows one to do is to determine the set of acceptable aggregator designs, which can be a very helpful piece of information for assessing the question of balancing volume and nuanced analysis of fewer pieces of evidence.

## 3.5 Potential objections to the proposed criterion

To conclude, I consider potential criticisms of my approach to assessing the optimal extent of automation of an evidence aggregator. On the one side of the spectrum, it might be objected that I rely too much on intuitive reasoning or expert judgement, when it would be better to appeal to objective assessments of the track record of evidence aggregators. On the opposite side, it might be objected that there is no plausible alternative to implicit expert reasoning when it comes to aggregating diverse evidence.

Two things can be said in response to the first objection. To begin with, in the case that one could objectively assess the track record of an automated evidence aggregator, the ability to perform a robustness analysis would still be an important consideration. This ability would allow comparing the performance of multiple versions of an aggregator and, hence, selecting the best aggregator. Second, it is not obvious of what an objective assessment of the track record of an aggregator would consist. A clear criterion for an aggregator having produced a verified result would be an instance in which the aggregator predicts an event and the event in fact takes place. However, the kinds of hypotheses considered in this chapter (e.g., 'treatment 1 is better than treatment 2 with respect to breast cancer rates') are generally not of

this kind. Medical hypotheses tend to be stochastic (and moreover, associated predictions often depend on a number of auxiliary conditions); thus, there is no definitive point in the future at which evidence would be received that would settle such matters.

The second objection is: The automation of evidence amalgamation is doomed to fail because it is inherently resistant to explicit algorithms. To start, one might argue that judgement is required at a very basic level, namely for the application of any (reliability) criterion to a particular experimental finding. Even if the reliability criterion is expressed in a functional form in terms of input variables and parameters, this objection goes, the value assigned to the input variables in any concrete case requires expert judgement. Stegenga (2014, p. 203) cites empirical studies showing that applying the same quality assessment tools to the same medical finding leads to widely diverging quality assessment of this medical evidence, because the finding itself is perceived differently by different investigators. Let me respond by pointing out that I accept the importance of expert judgement in the process of evidence aggregation. Yet, I wish to stress that there is no reason to deny that at least some of the reasoning involved in evidence aggregation/inference can be broken down and made explicit via an algorithm. Indeed, aggregators may be deemed more or less transparent depending on the extent to which they explicitly account for the all the steps underlying the final assessment of hypotheses.

In addition, one might argue that evidence amalgamation is resistant to explicit algorithms since every case of amalgamation is distinct, and, hence, an algorithm, which is necessarily more general, misses the distinct features of the particular case. I do not find this objection convincing. Surely, there is a sufficient degree of similarity between classes of evidence aggregation problems that allows formulating aggregation rules that can be encoded in an automated aggregator. The quality assessment tools mentioned by Stegenga (2014) are a case in point.

## 3.6 Concluding remarks

In this chapter, I engaged with Hunter and Williams' automated evidence aggregator to address my second research question: What is the value of derivational robustness analysis if not the entire relevant possibility space is covered by the available models? As it turned out, the structural reflections about evidence aggregation algorithms are the same as the one that can be asked about a set of models. I put forward a new criterion for assessing automated evidence aggregators. This criterion reveals a role of derivational robustness analysis even in cases where an epistemic agent knows that not the entire relevant possibility space has been covered in a robustness analysis. Derivational robustness analysis provides a structured way of establishing a preference relation over different algorithm designs by emphasising the amount of relevant possibility space covered. Due to the analogy between algorithms and models, this analysis shows that derivational robustness analysis provides a structured way of forming preference relations over different model structures, where the preference relation is based on the amount of the relevant possibility space that is covered by a model.

As a spin-off of my engagement with Hunter and Williams's proposal, I argued that where there are resource constraints on the aggregation process, one must also consider what balance between volume of evidence and accuracy in the treatment of individual evidence best facilitates inference. Again, concentrating on robustness analysis helps here, but there are further trade-offs between nuanced analysis of evidence and volume of evidence that must be taken into account. There is no requirement to aggregate the total evidence available if this would in fact reduce overall accuracy.

In the following chapter, I turn to policy-making in the economic domain. This enables me to examine further issues surrounding derivational robustness analysis, in particular the question of what to do if the robustness condition fails across a set of models. Before I address this question, I shall argue that economic policy-making reveals that one can distinguish an additional fruitful category of robustness analysis.

# Chapter 4

# Robustness in Economic Policy-Making: A New Characterisation of Toy Models

## 4.1 Introduction

Observing a group of children is interesting for many reasons. One is surely their ability to engage in all kinds of play. They spend hours building sand sculptures, throwing marbles, or pretending to be characters from the middle ages. A look into the psychological literature on child development reveals that these activities foster the development of important cognitive and emotional capabilities, such as creativity (Singer and Singer 1990), self-regulation (Hirsh-Pasek et al. 2009), and problem solving (Smith and Dutton 1979). These capabilities are particularly promoted by pretense-play, that is the intentional projection of mentally represented alternatives on a given situation in the spirit of play (Lilliard et al. 2013, p. 2).

Scientists in a variety of fields also engage in activities that have been characterised in play-like terms. Most notably, scientists often construct and explore *toy models* (Hartmann et al. 2016; Sugden 2000; Kuorikoski and Lehtinen 2009; Grüne-Yanoff 2009; Grüne-Yanoff 2013a). Such models play a central role in science, including in social sciences. For example, Schelling (1971)

introduced a toy model of racial segregation. In this model, cities are represented as two-dimensional grids, inhabited by stars and dots that move around to satisfy their preference not to be in the minority in their neighbourhood. The model is regarded as a paradigmatic case of a good social scientific model (Sugden 2000, p. 29).

Toy models are particularly widespread in economics.[1] Morgan (2013, p. 384) observes the following:

> To an outsider coming to the field of economics, one of the most striking things is the way that economists feel that they can express so much of what happens in the economy with their small worlds, within these little chunks of mathematics or puzzling diagrams. (...) Economic models have occasionally been referred to as 'toy models' (by both critics and users), conjuring up images of the scale models of farm animals and fire engines.

Indeed, economists describe their own practice in these play-like terms. Romer (1993, p. 66), a famous macroeconomist, explicitly uses the term 'toy model' and remarks that the "label is apt because a good theoretical model should be as easy to manipulate in one's head as the mental image of a child's toy". Jaeger (2009) reported the increased usage of the notion 'toy model' in economics, before dismissing it since it had not yet been defined clearly.

Economic toy models influence economic policy-making. Economic policy decisions are frequently justified with reference to a body of economic knowledge. As, for example Kocherlakota (2009, p. 19) states, it was the case that a significant part of policy during the recent financial crisis was not based on particular models but on a broader set of background beliefs, including verbal intuitions and crude correlations. Toy models shape these background beliefs. Akerlof's market for lemons model, which discuss in more detail below, illustrates this role of toy models in shaping relevant background beliefs. In 2001, Akerlof shared the Nobel Prize for economics with Joseph Stiglitz and Michael

[1] This statement should not be read as suggesting that state of the art economics is *mostly* a theoretical exercise. As Smith (2016) makes clear, economics has become more empirically orientated over the last five decades.

Spence, who were joint winners for their work on informational asymmetries and market behaviour. Akerlof's share of the prize was awarded largely for his paper "Market for Lemons", where he showed that in markets, in which the sellers know more about the product than buyers, low-quality products can squeeze out high-quality products (Akerlof 1970). This publication is not only one of the best-known papers in theoretical economics (Sugden 2000, p. 2), but the concept of informational asymmetries has influenced the regulation and design of many markets. The model has been influential despite the fact that, as will become clear below, it does not give precise quantitative predictions about the role of different types of informational asymmetries that can arise in markets.

In this chapter, I look closer at the notion of a toy model and provide a new characterisation of this term that enables me to address my first research question: Do measurement and derivational robustness analysis exhaust the set of useful types of robustness analysis? As it turns out, toy models can be characterised by a particular form of stability requirement which I will call *predictive stability*: abstract (or higher-level) properties, which are exemplified by toy model results, are robust under changes of target systems; or to put it differently, toy models come equipped with the hypothesis that their results are present across different target systems since the results do not depend on the lower-level configuration of these systems.

This chapter is structured as follows. First, I introduce five models which have been characterised as toy-like in the literature and which exhibit particularly interesting features (*Section 4.2*). Second, I show how existing characterisations of models fail to adequately delineate this set (*Section 4.3*). Third, I propose a new approach to characterising toy models. I take the five models as the starting point and identify three condition that jointly characterise them: i) a manipulability condition regarding the derivation of key model results, ii) a relation of multiple realisability between key model results and a set of diverse target systems, and iii) a non-representation relation between the basic entities and properties of a model and target systems (*Section 4.4*). I then highlight three corollaries of my explication (*Section 4.5*).

## 4.2 Five case studies

In this section, I introduce five models. These models are particularly interesting cases of toy models as they give a totally wrong-headed picture of reality. The examples stem from a variety of scientific disciplines: from physics (Kac ring), economics (Akerlof's market for lemons, Hotelling's model, DY model), and sociology (Schelling's checkerboard model). Although the focus of this and the next chapter is the epistemic value of toy models for economic policy-making, models from physics and sociology are also discussed, since the primary aim of this chapter is a conceptual clarification of the term 'toy model'.

My characterisation is based on these five models. These five cases, however, do not exhaust the models that are labelled 'toy' in scientific practice and by the philosophy of science literature. As will become clear from the discussion below, there is some disagreement about how the cases should be classified. To start, the term 'toy model' is used across disciplines to denote models that solely represent one causal (or explanatory) factor or mechanism in a target system. After I have presented my account of toy models, I am able to state precisely a crucial difference between these models and my five identified cases. Because of this difference, I suggest distinguishing between toy models and, what I call, one-factor models (see Section 4.5.1). The term 'toy model' is also used to denote models that are employed for exploring theories. With respect to this usage of term, I will suggest that one should group such models under the notion of probing models (see Section 4.5.2).

The main reason for these two, admittedly invasive, terminological moves is to achieve the clearest possible view of the set of 'toy-like' models. My account of toy models identifies a particularly interesting set of models; interesting, because they pose a particular challenge to learning with models, as I argue in Chapter 5.

### 4.2.1 Schelling's checkerboard model

Schelling (1971)'s checkerboard model is widely regarded as toy-like (see Hart-

mann et al. 2016; Thébault et al. 2016; Jebeile 2016; Casini 2014; Ylikoski and Aydinonat 2014; Grüne-Yanoff 2009; Grüne-Yanoff 2013a; Cartwright 2009; Sugden 2000). Schelling's model focuses on the "interactive dynamics of discriminatory individual choices" (Schelling 1971, p. 143). Discriminatory choices are those which reflect, consciously or unconsciously, identity features, such as sex, age, religion, or race (Schelling 1971, p. 144).

The key insight of the model is that one can end up with segregation even with very mild discriminatory preferences, such as the preference that one is not in a minority (Schelling 1971, p. 156). Although, Schelling applies the model to the segregation of black and white citizens in the cities of the United States, he envisages a much broader domain of application (see Schelling 1971, p. 144):

> The analysis, though, is so abstract that any twofold distinction could constitute an interpretation — whites and blacks, boys and girls, officers and enlisted men, students and faculty, teenagers and grown-ups. The only requirement of the analysis is that the distinction be twofold, exhaustive, and recognizable.

How does the model work? Schelling assumes that a population can be divided into two groups. Membership in a group is permanent. Everyone is assumed to care only about the composition of their neighbourhood in terms of members. Everyone has a particular location at every given moment in time and is, if not satisfied with the neighbourhood, capable of moving (Schelling 1971, p. 149). He introduces a one- and two-dimensional version of the model.

In the one-dimensional version of the model, a given finite number of stars and zeros are allocated in a random initial distribution on a straight line. For each element of the line a neighbourhood is defined that contains the four elements left and right of the particular star (or zero). Each element of the line is endowed with a preference over the composition of its neighbourhood: each element wants at least half of its neighbours to be like itself. Completing the model with a movement rule generates the dynamics of the model: If every element moves to the nearest spot that satisfies its preference and the order

of movement is from left to right, then, after a short sequence of movements, a segregated pattern emerges on the line (Schelling 1971, p. 151).

In the two-dimensional version of the model, the elements can move in a plane. This yields a checkerboard structure (see Figure 4.1). Assuming a suitable change in the neighbourhood definition (the neighbourhood includes the eight adjunct cells of a specific cell on the grid), the preferences defined over these neighbourhoods (a fixed percentage of members of the neighbourhood needs to be of the same colour), and the movement rule (the movement happens roughly from left to right areas whereas elements move to the closest spot determined by the number of squares they have to traverse horizontally and vertically), patterns of segregation emerge (Schelling 1971, p. 157). Both versions of the model do not presuppose any anticipation of the movements of other agents (Schelling 1971, p. 150).

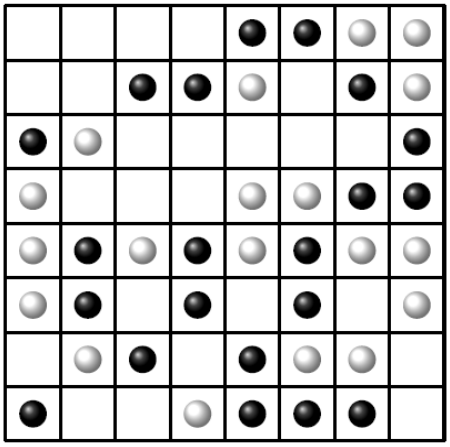

Figure 4.1: Two-dimensional version of Schelling's checkerboard model with a random initial distribution (Hartmann et al. 2016, Figure 1).

### 4.2.2 Akerlof's market for lemons

In his paper *Market for Lemons: Quality Uncertainty and the Market Mechanism*, Akerlof investigates the role of information in the working of markets. The model of a car market is widely regarded as toy-like (see Hartmann et al. 2016; Kuorikoski and Lehtinen 2009; Sugden 2000).

The key insight of the model is the negative role that differences in information between sellers and buyers of goods or services (informational asymmetries) can play. More precisely, markets can collapse entirely if the informational asymmetries are too large (Akerlof 1970, p. 488; Sugden 2000, p. 4-5). Although Akerlof discusses the car market, he has a broader domain of application in mind (Akerlof 1970, p. 488, 489). He makes this particularly clear in an autobiographical reflection:

> Indeed, I soon saw that asymmetric information was potentially an issue in any market where the quality of goods would be difficult to see by anything other than casual inspection. Rather than being a handful of markets, the exception rather than the rule, that seemed to me to include most markets. (Akerlof 2001, p. 4)

How does the model work? Akerlof assumes that there are only four types of cars: New cars vs. old cars and good vs. bad cars. Akerlof calls bad cars 'lemons' and good cars 'peaches'. Individuals in the market who buy a new car do not know its quality at the moment of purchase, but do know that with probability $q$ it is a peach and with probability $(1-q)$ it is a lemon. However, after having owned a specific car for a length of time, the car owner can form an idea about the quality of his or her car. In this situation, an asymmetry of information has developed between the car owner and the potential buyer (Akerlof 1970, p. 489). In this situation, the buyer of a used car is not willing to pay more than the average expected quality of a car. Owners of peaches are not willing to sell their cars on the market because they will not receive an adequate price for their above average quality car. Hence, the peaches are not marketed at all and one ends up with a market for lemons (Akerlof 1970, p. 490).

Akerlof complements this analysis with a case in which the quality of the product is not discrete (i.e., peaches vs. lemons) but continuous. In this case the conclusion is even starker: one observes a complete drying-up of the market, that is, no trade takes place. To show this, he introduces an explicit demand and supply function for the used car market and investigates the price assumption, under which trade happens. He distinguishes between two

types of traders: group one and two. They are endowed with the following utility functions (where $M$ is the consumption of goods other than cars, $x_i$ is the quality of the $i$-th car, and $n$ is the number of cars) (Akerlof 1970, p. 490):

$$U_1 = M + \sum_{i=1}^{n} x_i \tag{4.1}$$

$$U_2 = M + \sum_{i=1}^{n} (3/2)x_i \tag{4.2}$$

Both types of traders are assumed to be von Neumann-Morgenstern maximisers of expected utility. Group one has $N$ cars with uniformly distributed quality ($0 \leq x \leq 2$) and the price of $M$ is unity. These assumptions allow deriving the demand and supply of cars by group one and two, respectively. Adding up the demand for group one and two gives the following description of total demand $D$, where $Y_i$ is the income of group $i$ of traders and $\mu$ is the average quality of cars in the market (Akerlof 1970, p. 491):

$$D = \begin{cases} (Y_1 + Y_2)/p & \text{if } p < \mu \\ (Y_2)/p & \text{if } \mu < p < (3/2)\mu \\ 0 & \text{if } p > (3/2)\mu \end{cases} \tag{4.3}$$

Since with price $p$ the average quality of cars in the market is going to be $p/2$, none of the conditions set out in the equations (4.3) are fulfilled at any price level, and, hence no trade takes place (Akerlof 1970, p. 491).

To show that the insight of the model is also transferable to other market contexts, Akerlof (1970, p. 492) discusses the case of medical insurance. Older people have difficulty getting medical insurance since, as the premium increases with age, only those who are increasingly certain that they need the cover will insure themselves. This means that the average health level of the insured deteriorates as the price level rises, resulting in a situation in which no one can get insurance regardless of the price level.

### 4.2.3 DY model

The Dragulescu and Yakovenko (DY) model (Dragulescu and Yakovenko 2000) is a toy model from econophysics and has recently been picked up in the philosophical literature (see Hartmann et al. 2016 and Thébault et al. 2016). Econophysics is the relatively young field that applies models from physics to questions in economics; mostly methods of statistical mechanics are used to analyse financial market phenomena (Rickles 2007). The DY model is a model of monetary income distributions in societies. Monetary income distributions can be visualised by plotting the actual income per annum against the cumulative percentage of people earning this income (see Figure 4.2). The DY model treats agents as analogous to molecules and their interactions as analogous to molecular collisions in which kinetic energy is exchanged. These collisions lead to the observable income distributions. This contrasts with the neo-classical macroeconomic approach to explaining income distributions, which sees the interplay between technological development, education level of workers, and their marginal productivity as key drivers. In short, given that there is educational progress in a society, if there is a long period without significant technological change, then the education levels of lower-skilled workers catch up, their marginal productivity increases, and, hence, a more equal income distribution results. In contrast, if there is rapid technological change, then the education levels of lower-skilled workers will not catch up, a differential in marginal productivity compared to better educated workers remains, and, hence, a more unequal income distribution results (Thébault et al. 2016, p. 5-6).

The key insight of the DY model is that the bulk of the monetary income distribution is exponentially distributed. This observation can be made for a wide range of real income data (Thébault et al. 2016, p. 6).[2]

How does the model work?[3] The DY model assumes that there is a large

[2]To keep the exposition simple, I focus here on the income distribution's exponential bulk. As displayed in Figure 4.2, income distributions also show a power law tail. This second fact can be recovered by adding random savings to the DY model (Thébault et al. 2016, p. 3, 16).

[3]I rely on the exposition of Thébault et al. (2016), which parallels the one of Dragulescu and Yakovenko (2000) but is clearer.

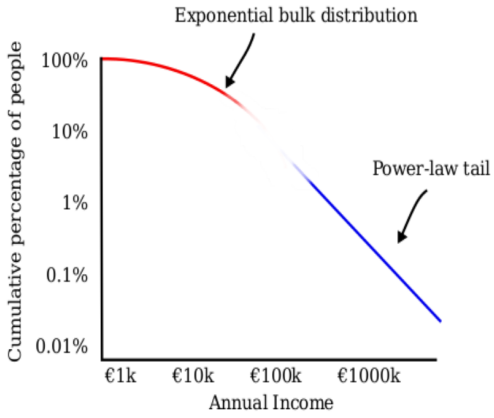


Figure 4.2: Representation of an income distribution (Thébault et al. 2016, Figure 1).

population of agents with zero intelligence (i.e., they have no capacity for anticipatory or strategic behaviour). At any given time $t$ an agent $i$ is associated with an amount of money $m_i$ that is non-negative. At any given time $t$, two individuals $i$ and $j$ are randomly selected from the population. Their interaction is modelled as a binary exchange of money. When they meet, their pre-interaction money $m_i$ and $m_j$, respectively, is pooled. From this pooled money a random fraction $\epsilon_{ij}$ is given to agent $i$ and the rest to agent $j$. During all the interactions the number of agents and the total amount of money stays the same. The post interaction money allocation looks like this (Thébault et al. 2016, p. 8):

$$m_i(t+1) = m_i(t) + \Delta m \tag{4.4}$$

$$m_j(t+1) = m_j(t) - \Delta m \tag{4.5}$$

where $\Delta m$ denotes the amount that is exchanged between the agents and can be expressed as:

$$\Delta m = \epsilon_{ij}(m_i(t) + m_j(t)) - m_i(t) \tag{4.6}$$

Given this set-up, a probability distribution over the monetary income of

the agents can be defined. To do this, bins of monetary amounts are delineated such that a bin has the width $a$.[4] Let $N_k$ denote the number of agents with income between $m_k$ and $m_k + a$. The probability of an agent falling into bin $k$ ($p(m_k)$) is $\frac{N_k}{N}$. Now, the fact that a specific number of agents is in bin $k$ can be realised in multiple ways (since only the number of agents is relevant and, hence, two sets with equal cardinalities representing different combinations of agents are not distinguished). Given this, one can express the number of distinct placements of agents that establish the same number of agents in bin $k$ as:

$$\Omega = \frac{N!}{k!(N-k)!} \tag{4.7}$$

Now, the DY model exploits the fact that the natural logarithm of (4.7) corresponds to the entropy measure in statistical mechanics.[5] Borrowing the fact from statistical mechanics that equilibria in thermodynamic systems are states in which entropy is maximised enables one to find the equilibrium probability distribution over the income bins by maximising $ln\Omega$ (see Dragulescu and Yakovenko 2000, p. 724). Performing this maximising procedure yields the following probability distribution (Thébault et al. 2016, p. 9):

$$P(m_k) = \frac{N_k}{N} = e^{-\frac{m_k - \mu}{T}} \tag{4.8}$$

with $T = \frac{M}{N}$ and $\mu = -Tln\frac{T}{a}$. The distribution $P(m_k)$ is an exponential distribution and has a bulk. Hence, the DY model reproduces one of the facts about monetary income distributions.

### 4.2.4 Hotelling's model

Hotelling's model (Hotelling 1929) is widely regarded as toy-like (see Reiss 2012; Alexandrova and Northcott 2013; Grüne-Yanoff 2013b; Hausman 2013; Mäki 2013; Rol 2013). Hotelling's model is used to study monopolistic com-

[4] I part here with the notation used by Thébault et al. (2016, p. 9) to simplify the exposition.

[5] This concept will be introduced in more detail in discussion of the Kac ring (see Section 4.2.5).

petition, that is, situations in which sellers of a product or service have a certain market power realised as price-setting capacity.

The key insight of the model is expressed as Hotelling's law of minimal differentiation (Reiss 2012, p. 44). The law states that competitors supply products or services that are very similar (Reiss 2012, p. 46). Hotelling (1929, p. 54, 57) observes this in many economic situations:

> It [the tendency to minimal differentiation] leads some factories to make cheap shoes for the poor and others to make expensive shoes for the rich, but all the shoes are too much alike. Our cities become uneconomically large and the business districts within them too concentrated.[6]

Hotelling also observes this phenomenon in many non-economic situations (Hotelling 1929, p. 54, 57):

> In politics it is strikingly exemplified. The competition for votes between the Republican and Democratic parties does not lead to a clear drawing of issues, and adoption of two strongly contrasted positions between which the voters might choose. Instead, each party strives to make its platform as much like the other's as possible (...) Methodist and Presbyterian churches are too much alike.

How does the model work? Assume that buyers of a commodity are uniformly distributed along a line with length $l$. At distances $a$ and $b$ from the endpoints of this line two producers $A$ and $B$ are located (see Figure 4.3).

Each buyer consumes one unit of the commodity per time interval and must transport the bought unit of commodity to his or her location with a cost of $c$ per unit distance. The production cost per unit commodity is zero for both producers. The consumer's preferences for producers are solely determined by the total of the price per unit commodity and the transportation costs.

[6]The normative judgements expressed here ("too much alike", "too concentrated") refer to the fact that the resulting equilibrium allocations are not socially optimal. I take this point up in my description of the details of the model.


Figure 4.3: Two producers are positioned along a continuum of buyers in the basic set-up of the Hotelling model (my diagram).

Hotelling denotes the prices of producers $A$ and $B$ as $p_1$ and $p_2$, respectively ($q_1$ and $q_2$ are the sold quantities at these prices) (Hotelling 1929, p. 45).

Given these assumptions, one can determine the equilibrium position of the two producers $A$ and $B$ along the line. To do this, consider first consumer $C$ who is indifferent between buying from producer $A$ or $B$ (see Figure 4.4).

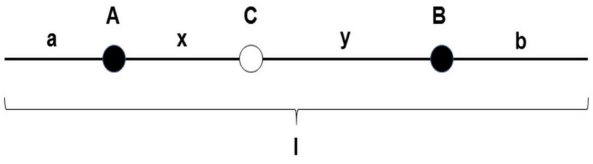


Figure 4.4: Consumer C is indifferent between producers A and B since the total costs of acquiring the goods either from A or B are identical (my diagram).

For this consumer, the costs associated with acquiring the unit of the commodity from either of the two producers need to be equal (where $x$ and $y$ is the segment of $l$ which connects $A$, or $B$ respectively, with the consumer of interest):

$$p_1 + cx = p_2 + cy \tag{4.9}$$

Combining this with the observation that $a + x + y + b = l$, one can

formulate the profits for producers $A$ and $B$ as (Hotelling 1929, p. 46):

$$\pi_1 = p_1 q_1 = p_1(a + x) = \frac{1}{2}(l + a - b)p_1 - \frac{p_1^2}{2c} + \frac{p_1 p_2}{2c} \tag{4.10}$$

$$\pi_2 = p_2 q_2 = p_2(b + y) = \frac{1}{2}(l - a + b)p_2 - \frac{p_2^2}{2c} + \frac{p_1 p_2}{2c} \tag{4.11}$$

Differentiating both equations (4.10) and (4.11) with respect to the prices $p_1$ and $p_2$ and setting the resulting equations to zero yields the following optimal price-setting behaviour for the two producers (Hotelling 1929, p. 46):

$$p_1 = c(l + \frac{a - b}{3}) \tag{4.12}$$

$$p_2 = c(l - \frac{a - b}{3}) \tag{4.13}$$

Given the equations for the optimal quantities of $A$ and $B$ (following from the same optimality condition), one can express the profits of $A$ and $B$ as (Hotelling 1929, p. 50):

$$\pi_1 = \frac{c}{2}(l + \frac{a - b}{3})^2 \tag{4.14}$$

$$\pi_2 = \frac{c}{2}(l - \frac{a - b}{3})^2 \tag{4.15}$$

Now, assume that the position of the producers on the line is not fixed. $A$ tries to maximise profits by increasing $a$. $B$ tries to maximise profits by increasing $b$. Hence, if they can move, the two producers move as closely together as possible without occupying exactly the same spot on the line (this would end in a price war that reduces the profits for both) (Hotelling 1929, p. 52). Note, finally, that the resulting allocation is not socially optimal. If $A$ and $B$ were to end up at the 1/4 and 3/4 point of the line, they would also split the profits but the customers would have to travel less (Reiss 2012, p. 46).

### 4.2.5 Kac ring

The Kac ring (Kac 1959) is a toy model from statistical mechanics (see Jebeile 2016; Gottwald and Oliver 2009, p. 614). To be able to express its key insight, I must present some background regarding the second law of thermodynamics and time-reversible physical processes.

Thermodynamics is the field of physics that describes phenomena from a macro-perspective, such as the temperature, volume, or pressure of a gas contained in a box. Statistical mechanics adopts a micro-perspective on the same phenomena, that is, on the micro-constituents (such as molecules of the gas) of the macro-phenomena and their dynamical laws. Statistical mechanics aims more specifically to account for the behaviour of macro-phenomena in terms of their micro-constituents (Frigg 2010, p. 1).

One observable fact is that many processes are unidirectional: one sees coffee being spilled over a laptop but not the coffee flying back into the mug; one sees gin and tonic mixed in a drink but not the un-mixing of gin and tonic; one sees a window being shattered by a rock but not the spontaneous assembly of the glass pieces into a window (Frigg 2010, p. 2). This fact is captured by the *second law of thermodynamics*. This law states, roughly, that transitions from an equilibrium (such as the state of a well-mixed gin and tonic) to a non-equilibrium state (such as the un-mixed gin and tonic) cannot occur in isolated systems (Frigg 2010, p. 2), where isolated means that the system does not exchange energy in the form of heat or work with any other system (Reif 1985, p. 91). A key question of statistical mechanics is how one can account for this second law of thermodynamics from the micro-perspective.

In fact, if one attempts to derive the second law of thermodynamics from assumptions in statistical mechanics, one quickly encounters a problem, termed the *Loschmidt reversibility paradox* (Frigg 2010, p. 12). Consider a container with two compartments that are connected by a shutter. At the initial point in time ($t = 0$), a gas is contained in the left compartment and the shutter is closed. Then, the shutter is opened, the gas spreads out, and after some time it is distributed uniformly across the two compartments ($t = 1$). The gas has reached its equilibrium state. Now, assume that all the veloc-

ity vectors of the gas molecules are reversed, then the gas molecules would travel along their initial trajectories back into the left compartment of the box ($t = 2$). Hence, it seems possible to violate the second law of thermodynamics without violating the dynamical laws constituting the system under consideration (Henderson 2014, p. 90).[7]

The key insight from the Kac ring is that a system with a deterministic, time reversible microdynamic and an ergodic decomposition shows an *equilibrium-like behaviour*. An ergodic decomposition is the decomposition of a system into its ergodic components. A component is ergodic if it is the case that, on average, the time it spends in a subset of the phase space is proportional to the portion of the phase space occupied by that subset (Frigg 2010, p. 10). Put informally, the evolution of a system is equilibrium-like if the systems' entropy is close to its maximum value, from which it exhibits frequent small fluctuations, and rarer large fluctuations (Lavis 2008, p. 684). Figure 4.5 illustrates the difference between equilibrium behaviour and equilibrium-like behaviour.

The insight from the Kac ring can be generalized. As Werndl and Frigg (2016, p. 9) have shown, one can prove an ergodic decomposition theorem stating that every measure-preserving dynamical system has an ergodic decomposition, and, consequently, shows equilibrium-like behaviour.[8] This theorem applies widely since most of the dynamical systems surrounding us, even those with a very complex dynamic, are measure-preserving.

How does the model work? The Kac ring[9] consists of $n$ equidistantly distributed sites on a circle. On each site, there is either a white or a black ball. $m$ of the intervals between the sites are marked and form a set $S$. The balls move counter-clockwise to their nearest site. If a ball is before a marked

[7]The notion of time reversibility is used in a non-technical way here. For what follows, this account of time reversibility suffices: "A physical law is time reversal invariant if whenever a motion is allowed by the law, there is a nemesis 'time-reversed' motion that is also allowed by the law, corresponding roughly to what one would see if a film of the original motion were played in reverse" (Roberts 2013, p. 1113).

[8]To be precise, Werndl and Frigg (2016) prove this claim under the assumption of their long-run fraction of time definition of Boltzmannian equilibrium. See Werndl and Frigg (2015, p. 25) for the introduction of this equilibrium concept.

[9]I follow the discussion of Bricmont (1996, p. 41), which follows Kac (1959)'s exposition very closely but is clearer.

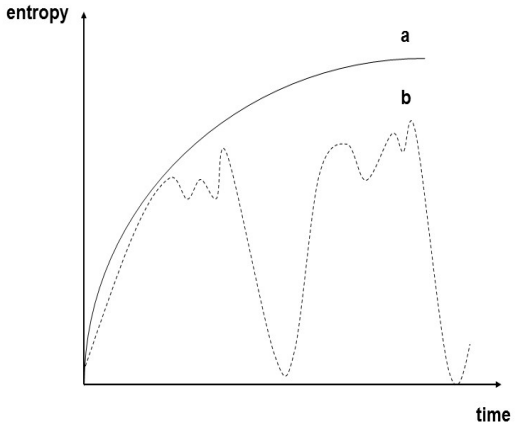


Figure 4.5: Level of entropy of two systems over time. System (a) shows behaviour in accordance with the second law of thermodynamics, that is, an increase of entropy over the observed time span. System (b) shows equilibrium-like behaviour (my diagram).

interval, then it changes colour upon passing it. If a ball is in front of an unmarked interval, then it keeps its colour in this movement step (Bricmont 1996, p. 38). A micro-state of the Kac ring is a ring with a distribution of black and white balls and a distribution of marked intervals (see for an example Figure 4.6). A macro-state of the Kac ring is defined in terms of the number of white ($N_w(t)$) and black balls ($N_b(t)$). A macro-state (e.g., 5 white and 5 black balls) can be instantiated by many micro-states (e.g., 5 black and 5 white balls placed alternately on the sites with 5 marked intervals).

The ring displays a time-reversible micro-dynamic (Bricmont 1996, p. 38). If after $t$ time steps one reverses the movement of the balls (from counter-clockwise to clockwise movement), then after $t$ time steps one returns to the original state. This micro-dynamic is furthermore completely deterministic (Bricmont 1996, p. 38).

The equilibrium state of the Kac ring is where there are an equal number of black and white balls. How close one is to this equilibrium state can be

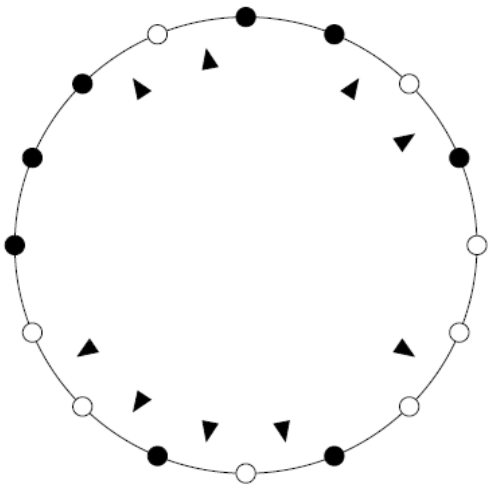

Figure 4.6: A Kac ring with 16 lattice sites and 9 markers (Gottwald and Oliver 2009, Figure 1).

expressed as the greyness of the ring, that is, the difference between the (normalized) number of white and black balls:

$$\frac{N_w(t) - N_b(t)}{n} \tag{4.16}$$

Now, one can define the Boltzmann entropy for the system of the Kac ring. The Boltzmann entropy of a macro-state is defined as the logarithm of a measure of the micro-states that realise this particular macro-state. For the Kac ring it is natural to define the entropy measure as:

$$S_b = ln\binom{n}{N_w(t)} \tag{4.17}$$

This entropy value is maximal for the states of equipartition between black and white balls (Bricmont 1996, p. 40).

This deterministic, time-reversible micro-dynamic shows an approach to equilibrium. If the ring contains an infinite number of sites, then there is a thermodynamic approach to equilibrium and, hence, irreversibility on the macro-level. For, if the ring starts in an initial configuration of uneven black and white balls and randomly distributed markers, it is much more likely

that the system evolves towards an equal distribution of black and white balls rather than a highly unequal distribution. However, if the ring contains a finite number of sites, then there needs to be reversibility also on the macro-level, since the ring returns to the initial configuration after a certain amount of time. In this scenario, however, the ring still shows equilibrium-like behaviour, since it exhibits an entropy profile for which it is the case that the entropy value of the system is close to the maximum value for most of the time considered.

## 4.3 Existing accounts of toy models

In this section, I review five existing accounts of toy models. I introduce these accounts in turn and argue that they highlight, individually, important features of the five cases introduced in the previous section. However, I go on to claim that one needs to piece these features together to achieve a satisfactory account of my cases. Before discussing the accounts, I add some preliminaries on the project of characterising toy models. These serve as an overreaching framework for discussing the existing accounts in the literature.

### 4.3.1 Preliminaries: Carnap on explication

According to Carnap, concepts must be made more precise to serve their role in scientific theories. The method he proposes is *explication*. Carnap (1962, p. 3) introduces this method as follows (italics in the original):

> By the procedure of *explication* we mean the transformation of an inexact, prescientific concept, the *explicandum*, into a new exact concept, the *explicatum* (...) The explicandum may belong to everyday language or to a previous stage in the development of a scientific language. The explicatum must be given by explicit rules for its use, for example, by a definition which incorporates it into a well-constructed system of scientific either logicomathematical or empirical concepts.

Explication, according to Carnap, can be viewed as a method of re-engineering concepts with the aim of advancing scientific theorising (Brun 2015, p. 1). One of Carnap's leading examples is the explication of the concept *fish* by the scientific concept *piscis*. 'Piscis' is defined as animals that live in water, are cold-blooded vertebrates, and have gills throughout life.[10] Let me illuminate the procedure of explication with reference to this example. First, the explicandum should be clear. This is not a trivial point. The whole project of explication gets off the ground because the explicandum is in a sense too imprecise for scientific usage. Hence, making the explicandum more precise does not mean that an exact definition should be given (see Carnap 1962, p. 4). Instead, the explicandum should be identified as clearly as possible. One way of doing this is to present cases to which the explicandum clearly applies or does not apply (Brun 2015, p. 4). With respect to the concept 'fish', one can clearly identify animals to which the concept applies and to which it does not apply. Second, an explicatum should be given. The explicatum is stated in a target system of concepts, that is, the concepts used in the domain of scientific theorising of interest (Brun 2015, p. 6). Usually a definition of the explicatum in terms of necessary and sufficient conditions is provided. Importantly, the explicatum and explicandum can be the same word (Brun 2015, p. 5). In the example of the concept fish, the terms 'vertebrate' and 'gill' are biological terms. Third, the explicandum is replaced by the explicatum and the adequacy of the explicatum is assessed. Carnap proposes four criteria that an adequate explication of a concept must meet (Carnap 1962, p. 7): i) The explicatum must be similar to the explicandum such that in most cases where the explicandum applies the explicatum applies as well. The concept 'piscis' applies to most things that in everyday language are classified as fish but not to whales; ii) the explicatum has to be exact, that is, the explicatum should be introduced in a well-connected system of scientific concepts; iii) the explicatum should be fruitful, that is, it should figure in many true generalizations (such as laws). The concept piscis

[10]Carnap makes it clear that the explicatum must not be introduced by a definition. Other methods for concept introduction can be used as well, for example, postulates or reduction sentences. Carnap introduced the latter to overcome the problem of characterising dispositional terms in observational terms (Brun 2015, p. 7).

figures in more biological generalisations than the pre-scientific concept fish; (iv) the explicatum should be as simple as possible. This requirement is best understood as a tie-breaker criterion: if two explicata satisfy criteria (i)-(iii) equally well, then the simpler one should be chosen (Brun 2015, p. 11).

I proceed in the same way in my characterisation of the term toy model. First, I identify a set of, in my view relevant, examples of toy models (see Section 4.2). Second, I introduce an explicatum of the term toy model in a step-wise fashion in the form of a definition (Sections 4.4.2-4.4.4). Third, I show that the explicatum meets the adequacy conditions set out by Carnap (Section 4.4.5).

### 4.3.2 Hartmann et al.'s account of toy models

In their recent article *Understanding (with) toy models*, Hartmann et al. (2016) suggest both a characterisation of the notion 'toy model' and an account of how one can gain understanding with toy models. They restrict their analysis to mathematical toy models in contrast to material toy models, which are physical entities (Hartmann et al. 2016, FN 3).

Toy models are characterised by Hartmann et al. (2016) by three features: First, toy models are strongly idealised in the sense that they contain Aristotelian as well as Galilean idealisations. Aristotelian idealisations strip away some features of a target system, for example, the colour of the block sliding down an inclined plane, or express the assumption that some causal factor is absent. Galilean idealisations, in contrast, are descriptions that distort the causal (or explanatory relevant) factors of a target system, for example, the assumption that the surface area indeed affects the time it takes a block to slide down an inclined plane (Hartmann et al. 2016, p. 6-7). Second, toy models are extremely simple as they represent only a small number of causal factors (or more generally explanatory factors). Third, toy models are target-directed models, that is, the models are built and evaluated in relation to (types of) target systems (Hartmann et al. 2016, p. 3). In passing, Hartmann et al. (2016, p. 3) also mention another feature of toy models; These models can be grasped with cognitive ease by people trained in a particular

field. Moreover, they do not think that there is a sharp distinction between toy and non-toy models. Instead, they claim that there is a continuum of models with respect to degree of simplicity and, independently, to degree of idealisation. Despite the vagueness of the concept, they think that there are clear cut examples of toy models (Hartmann et al. 2016, p. 3).

With these preliminaries in place, they introduce a distinction between embedded and autonomous toy models. A toy model is *embedded* if and only if it is a model of an empirically well-confirmed framework theory (Hartmann et al. 2016, p. 5). A framework theory is a set of uninterpreted sentences. Models of a framework theory are structures in which the sentences of the framework theory (e.g., the theory's abstract calculus and its laws) are true (Hartmann et al. 2016, p. 6). A framework theory is empirically well confirmed if the central predictions of the theory are accurate. Hartmann et al. (2016, p. 6) mention Newtonian Mechanics and Quantum Mechanics as empirically well-confirmed framework theories. The "Sun-plus-one-planet model" of the solar system is an example of an embedded toy model. It is a simple, highly idealised model in which the claims of Newtonian mechanics are true (Hartmann et al. 2016, p. 7). A toy model is *autonomous* if it is not a model of a well-confirmed framework theory. For example, there is no empirically well-confirmed framework theory of which Schelling's checkerboard model is an instantiation (Hartmann et al. 2016, p. 8).

How can one gain understanding with toy models? Hartmann et al. (2016, p. 19) suggest two ways of gaining understanding with toy models. They claim that an individual scientist $S$ understands a phenomenon $P$ via model $M$ in context $C$ if one of the following conditions holds:

- $S$ has how-actually understanding of phenomenon $P$ via model $M$ in context $C$ if model $M$ provides a how-actually explanation of $P$ and $S$ grasps $M$.

- $S$ has how-possibly understanding of phenomenon $P$ via model $M$ in context $C$ if model $M$ provides a how-possibly explanation of $P$ and $S$ grasps $M$.

An embedded toy model generates how-actually explanations only if further conditions are met (Hartmann et al. 2016, p. 19-20). First, the embedding framework theory permits an interpretation and justification of the idealisations of the model. Second, this interpretation and justification needs to be compatible with the veridicality condition, which states that explanatory assumptions are required to be true or approximately true (Hartmann et al. 2016, p. 15). Hartmann et al. (2016, p. 20) claim that key examples of autonomous toy models do not provide how-actually explanations; They are best interpreted as providing how-possibly explanations. Schelling's checkerboard model can be understood as providing an explanation of how it is possible that racial segregation patterns occur (Hartmann et al. 2016, p. 26).

Their account identifies key features of my five cases. To start, all five models that I introduced in the previous section can be explored with ease by a person trained in the respective field. This ease of manipulability seems indeed to be linked to a cognitive ease with which central derivations of toy models can be grasped by a person trained in the field. Furthermore, the five models contain multiple Aristotelian and Galilean idealisations. The Schelling checkerboard model omits social factors that are relevant for real world segregation patterns. The DY models omits the ability to save a certain amount of money, to invest in multiple financial instruments, and to entertain financial transactions with multiple persons at a given time. Akerlof's market for lemons and Hotelling's model distort the calculation and memory capability of real world agents. The Kac ring distorts the interaction structure of particles in a gas by assuming that the particles move counter-clockwise on a ring. In addition, the five models are evaluated in relation to real world targets. For example, Akerlof discusses the implication of his used car model for the insurance market and the money lending market. Hotelling himself foresaw the applicability of the principle of minimal differentiation to the interaction between political parties. Finally, the five toy models are simple in the sense that they do not contain a large number of causal (or explanatory) factors.

However, there are, in my view, two aspects of the five examples that are not captured by Hartmann et al. (2016)'s account. First, on the level of

model description, these cases give a particularly wrong-headed picture of the target systems of interest. Take Schelling's checkerboard for example. The description of a city as consisting of squares inhibited by dots of two different colours is such a misrepresentation that it seems to be far-fetched to say that Schelling had to discover (or learn) that this model is a misdescription of the situation. The same holds for the Kac ring. To describe a gas as a ring populated by balls of different colours and markers is to give a completely wrong-headed picture of a gas. To strengthen this point, it does not seem to be the case that these features of the models can be linked to features in a target system via a process of de-idealisation. Since there is nothing that corresponds to the ring structure of the Kac model in a gas, it is unclear how one can arrive at a correct description of a gas by adding features to or changing features of the Kac ring. I will spell out these observations in more detail in Section 4.4. Second, Hartmann et al. (2016) do not incorporate the fact that the five toy models are applied to a variety of target systems. Take Schelling's checkerboard model as an example. Schelling himself envisages that the model can be applied to any target systems that exhibits a distinction that is twofold, exhaustive and recognisable (e.g., students and faculty or teenagers and grown ups). Hence, the results of Schelling's model are claimed to be stable across a variety of target systems.

### 4.3.3 Strevens and Weisberg's account of minimal models

Weisberg (2007) and Strevens (2008) introduce the notion of 'minimal models'. Although they do not refer to minimal models as toy models, one can ask whether the criteria they suggest can be used to characterise my five cases.

According to Strevens (2008, p. 315-329) minimal models are idealised models that truthfully represent two kinds of facts: to start, facts about a minimal set of explanatory relevant factors including true causal laws and statements about initial conditions for the target system under consideration; furthermore, the fact that some factors are not explanatory relevant. To put it differently, idealising assumptions refer to explanatory irrelevant factors.

Weisberg (2007, p. 642) agrees with this characterisation of minimal models. According to him, "minimalist idealization is the practice of constructing and studying theoretical models that include only the core causal factors which give rise to a phenomenon." To put it more directly, a minimalist model contains only those factors that make a difference to the occurrence and essential character of a phenomenon of interest.[11]

Strevens (2004) offers the *kairetic account of scientific explanation* to make room for the explanatory power of minimal models. According to Strevens, minimal models provide a causal explanation, that is, a story about why the phenomenon of interest occurred. However, for this causal story only those factors are relevant that make a difference for the occurrence of the phenomenon of interest. Strevens defines making a difference as "causal entailment", which is defined as logical entailment in a causal model. Accordingly, a causal factor makes a difference only if one removes the factor from the causal model, the model then no longer entails the phenomenon's occurrence. Since minimal idealisations introduce false but not difference-making factors, minimal models can provide explanations according to the kairetic account.

Does Strevens and Weisberg's characterisation of minimal models fit my five cases? I do not think so. Take Schelling's checkerboard model for example. As Hartmann et al. (2016, p. 23) point out, if Schelling's model were a minimal model, then all its idealisation would need to refer to explanatory irrelevant factors. Only in this circumstance can the model be said to capture the key difference-making factors for segregation phenomena. This is, however, not the case. The claim that economic factors and other social factors do not have an influence on real world segregation patterns is wrong. The same can be said for the DY-model. The assumption that agents are not able to make any conscious investment and saving decisions is not explanatory irrelevant

[11] Weisberg (2007, p. 645-646) discusses in addition multiple-model idealisations that are similar to minimal model idealisations in the sense that they are not justified by reference to de-idealisation. They are, however, distinct to minimal model idealisations since it is not assumed that there is a single best model of the phenomenon. Gibbard and Varian (1978)'s notion of a *caricature* shows striking similarities to Weisberg's idea here. Gibbard and Varian define caricatures as models involving deliberate distortions of aspects of the target system where these distortions isolate effects or are used to test the robustness of a model claim (Gibbard and Varian 1978, p. 676).

for real world income distributions. Finally, regarding Akerlof's market for lemons, the cognitive capabilities of agents and their price-setting behaviour are explanatory relevant factors.

### 4.3.4 Grüne-Yanoff's account of minimal models

According to Grüne-Yanoff (2009, p. 83), a minimal model is a model that lacks world-linking properties. To be more precise, such models lack "any similarity, isomorphism, or resemblance relation to the world" and are assumed to "be unconstrained by natural laws or structural identity, and do not isolate any real factors".[12]

The upshot of Grüne-Yanoff (2009)'s discussion is the claim that minimal models in economics can perform their surrogative function despite the fact that the models do not have world-linking properties. How is this supposed to work? Grüne-Yanoff (2009, p. 81, 85) argues that one can learn from minimal models because results derived with the help of these models can affect one's confidence about impossibility hypotheses regarding the world in a justified way. Take Schelling's model as an example. One might hold the belief that racial segregation in cities cannot occur without strict discriminatory preferences. The model derivation of segregation patterns based on the assumption of only very mild racial preferences should affect our credence in the aforementioned impossibility claim. Grüne-Yanoff (2009, p. 94) adds the qualification that minimal models need to be credible (or plausible) to perform this epistemic function. He defines this notion of model plausibility in a purely model-internal way. According to him

> (...) judging a model to be credible is a consequence of what scientists do with models: they imagine a world that the model describes, they manipulate that situation in various ways, and they investigate that world's internal coherence and its coherence

[12] In the later Grüne-Yanoff (2013a), he drops the notion of 'minimal model' and speaks instead, with reference to the same and very similar examples, of non-representational models. He does, however, not offer a specification of what he means by non-representational models that goes beyond the explicit account given of minimal models.

> with our intuitions. Crucially, these intuitions often do not exist independently of the imagined world (...) credibility judgements about economics are often elicited solely through consideration of imaginary worlds. (Grüne-Yanoff 2009, p. 94)

Grüne-Yanoff's characterisation highlights one feature shared by my five cases: in an intuitive sense, there is a lack of world-linking properties between these models and their targets. Cities are not checkerboards, human beings do not collide to exchange money, and gases are not rings with markers. World linking properties seem to be absent on the level of the basic entities and properties of these models. As I specify below, the basic entities and properties of a model are those that are specified in the model description. To put it differently, they are the scaffolding of the model world. Model results are those statements that are derived from the model description. There are, however, multiple such world-linking properties to be observed in my five cases. These are in particular relations of similarity. Take the Kac ring for example. The model shows an equilibrium-like behaviour. This approach to equilibrium can be seen in many processes surrounding us. Market failures of the nature described in Akerlof's model (i.e., no transactions despite the fact that they could improve the buyers and sellers' utility) can be observed in many real world markets. The resulting income distribution in the DY model shows a striking similarity with the exponential income distribution in many developed economies. Hence, on the level of model results there are world-linking properties. One needs an account of models that differentiates in a more fine-grained manner between levels of model descriptions. In addition, Grüne-Yanoff's account also does not incorporate the observation, already mentioned in relation to Hartman et al.'s account, that the results from the toy modes introduced earlier are said to be stable across a variety of target systems.

### 4.3.5 Batterman and Rice's account of minimal models

According to Battermann and Rice (2014, p. 349), a minimal model is a caricature of a target system. It is a caricature in the sense that it "really

looks nothing like any system it is supposed to represent" (Battermann and Rice 2014, p. 350). These models explain "universal patterns across diverse real systems" (Battermann and Rice 2014, p. 350). By universal patterns, Batterman and Rice refer to macroscale behaviour. By diverse real systems, they mean systems that differ in their constitution on the microscale level. They discuss the Lattice Gas Automaton (LGA) as an example of a minimal model. The LGA describes a fluid as a set of particles confined to move on a hexagonal lattice. Adding some movement rules (involving a specification of what happens if particles collide) yields patterns that can be observed in fluid flow (Battermann and Rice 2014, p. 358).

How can one gain understanding with minimal models? Battermann and Rice (2014) argue that minimal models can provide *minimal model explanations*. Minimal models do not explain in virtue of sharing features with target systems of interest. Rather, the connecting element between minimal models and target systems is the fact that they belong to the same universality class (Battermann and Rice 2014, p. 350). Universality is an expression for the fact that many systems composed of different components show the same behaviour at higher scales (Battermann and Rice 2014, p. 357). One can show that a model and the target system it is supposed to describe belong to the same universality class by establishing that differences on the micro-level between the systems are irrelevant. This amounts to telling as to why very many features are irrelevant to the phenomenon that one seeks to explain (Battermann and Rice 2014, p. 361). To be more precise, a minimal model explanation is successful if it answers three questions (Battermann and Rice 2014, p. 361):

1. Why are the common features among the systems necessary for the phenomenon to occur?

2. Why are the remaining heterogeneous details (those left out of or misrepresented by the model) irrelevant for the occurrence of the phenomenon?

3. Why do very different target systems share common features?

Batterman and Rice suggest that question (2) can be answered by employing a mathematical technique called the renormalisation group. Without going deeply into technical details, the idea is as follows. One starts with a set of possible systems (e.g., possible fluids, actual fluids, a solid). One then applies an averaging technique to eliminate some degrees of freedom of this set of possible systems. One then rescales this set to generate a new set of possible systems. Examining this new set of possible systems allows identifying those systems whose behaviour depicted in a phase space flows to the same fixed point. These systems belong to the same universality class. One also learns the common features from this procedure. Furthermore, one sees why the systems share these common features as a consequence of the re-normalisation procedure, and, hence, one receives an answer to Question (3). Given the answers for Questions (2) and (3), one can answer Question (1) (Battermann and Rice 2014, p. 362-363).

Batterman and Rice's account highlights key features of my five cases. In particular, it brings out two features that have figured prominently in the discussion so far. First, these models seem to describe a behaviour that is shared across different target systems and that is best described on a macroscopic level (in contrast to the level of components of these different systems). Second, these models fundamentally mischaracterise their target systems.

However, there are aspects of Battermann and Rice (2014)'s account of minimal models that limit the applicability of their characterisation to my five cases. First, they provide what I will call a functional characterisation of minimal models.[13] They characterise this model class with the help of the explanatory function it performs. This makes it unclear how models that are similar to minimal models on the surface (i.e., they also look nothing like their target systems), but for which no satisfactory answers to the three questions can be provided, should be characterised. In a recent paper, Knuuttila and Loettgers (2016) argue that for social scientific models the group renormalization technique cannot be successfully applied. Extending their argument makes clear that Schelling's checkerboard model, Hotellings' model and Akerlof's market for lemons might also fall into this category. Second,

[13]I introduce this term in more detail in Section 4.4.1.

Battermann and Rice (2014) do not spell out their claim that a minimal model "really looks nothing like any system it is supposed to represent" in detail. What representational relation is in play here? Which level of description is supposed to represent and which one is not?

### 4.3.6 Bokulich's account of fictional models

Bokulich (2011, 2012) introduced the notion of 'fictional models'. Although she does not refer to fictional models as toy models, one can ask whether the criteria she suggests can be used to characterise my five cases.

Bokulich does not give an explicit definition of fictional models (or 'fictions' for short). However, she describes various characteristics of fictions and introduces illuminating examples. To start, a fiction is known to be false (Bokulich 2012, p. 725). Fictions are not idealisations since they are not related to target systems via a smooth process of de-idealisation (Bokulich 2011, p. 40). Bokulich discusses Bohr's model of the hydrogen atom as a model that contains fictional elements. In particular, there are no orbits along which electrons circumvent the nucleus of an atom. An electron should be better conceived as a cloud of probability density around a nucleus (Bokulich 2011, p. 42). One cannot recover the modern quantum mechanical picture of the hydrogen atom via a process of de-idealisation starting from Bohr's model. Adding additional factors or tweaking the parameter of Bohr's model does not do the trick. The two models of the hydrogen atom are in this sense not related in a continuous way with each other (Bokulich 2011, p. 43).

Can fictions deliver explanations? Bokulich thinks so. To state precisely how fictions can yield genuine explanations, she introduces a framework for model-based explanations. Model-based explanations are characterised by three features. First, such an explanation makes essential reference to a scientific model that involves a certain degree of idealisation and/or fictionalisation (Bokulich 2011, p. 38). Second, the model needs to correctly capture the counterfactual dependence in a target system. This means that the model needs to answer correctly a wide range of "what-if-things-had-been-different"-questions. Departing from Woodward (2003), Bokulich does not construe this

counterfactual dependence along interventionist lines and allows for counterfactual dependence that is not causal in nature (Bokulich 2011, p. 39). Third, there must be a justificatory step that consists in stating what the domain of applicability of the model is and in showing that the phenomenon that is to be explained falls into this domain (Bokulich 2011, p. 39). This justification can be given top-down by a theory that delineates the domain of applicability of a particular model (Bokulich 2011, p. 39).

It is this justificatory step that, according to Bokulich, allows one to distinguish between explanatory and non-explanatory fictions (Bokulich 2011, p. 39, FN 11; Bokulich 2012, p. 734). Explanatory fictions are representationally adequate. Representational adequacy should not be understood as the accurate description of a target phenomenon (fictions are false hypotheses about the target system) or as " saving the phenomena" (this is also accomplished by non-explanatory, phenomenological models) (Bokulich 2012, p. 734). Rather, the representational adequacy relation is dependent on the target system of interest, the epistemic aim one pursues with the modelling project, and the scientific community in which one is embedded in (Bokulich 2012, p. 734-735). Bohr's orbits are an explanatory fiction, since one can justify their representational adequacy, that is, semi-classical mechanics states that the orbits approximate quantum mechanical effects to a sufficient degree (Bokulich 2011, p. 44). The Ptolemaic epicycles are non-explanatory fictions since the present scientific community cannot provide a justification for their representational adequacy (Bokulich 2012, p. 727).

Bokulich's account of fictions helps to characterise key features of my cases. First, these models contain elements that are literally false statements about the target system of interest. More precisely, some of these fictional elements (such as the ring structure in Kac's model) cannot be de-idealised in a straightforward way. Second, nevertheless, these models are phenomenologically adequate in the sense that they reproduce patterns of data. However, Bokulich does not address the ease of manipulability of the five cases. The cognitive ease with which the models can be manipulated and results grasped, however, seems to be a distinctive feature of this model type (see Hartmann et al. 2016, p. 3). In addition, the stability of the model results across dif-

ferent target systems is not part of Bokulich's account. Finally, I think that Bokulich's account is somewhat unclear regarding the representational relation between fictions and real world targets. Bokulich (2012, p. 734) states the following:

> Where I want to part company with traditional philosophical accounts, however, is on the question of whether truth or existence is even a necessary condition for explanatory relevance. In particular, I want to argue that fictions can be explanatory relevant. That is, truth or existence is not a necessary condition for an item to be admitted to the scientists' explanatory store. However, such a position threatens to trivialize scientific explanation without some principled way of determining which fictions are to be counted as explanatory.

So, Bokulich suggests that there can be fictions that are genuinely explanatory despite the fact that they posit entities or properties that do not exist. She qualifies this claim by saying that these non-existent properties or entities are only explanatory relevant if there is some principled way of justifying them. However, this justification, as discussed above, is relative to a scientific community's standards of relevance. It is not clear to me to what extent these standards of relevance can be spelled out in representational terms or whether she has non-representational ways of achieving this justification in mind. To put it differently, what exactly is the representational status of those parts of models that are fictional in nature? I can relate my uncertainty here back to her leading example: Bohr's model of the hydrogen atom. In a sense, scientists believed that electrons circumvent the nucleus in classical orbits. A lot of experimental and theoretical work had to be done before it became clear that the orbits were a false representation of the electron's behaviour. In contrast, Kac proposing his model of the gas or Schelling presenting his model of segregation knew from the beginning that these models give a completely wrong-headed picture of their target system. This fact seems to be peculiar to my five cases. How can this observation be spelled out?

## 4.4 A new attempt at characterising toy models

The previous section should have made clear that none of the discussed accounts of models delivers, in itself, a satisfactory characterisation of my five cases. In particular, the accounts – beside Batterman and Rice's proposal – do not pay attention to the stability of the model results across different (types of) target systems. In this section, I provide a single, comprehensive account of my cases and suggest reserving the term 'toy model' for this model class.

### 4.4.1 Preliminaries: Functional vs. intrinsic characterisations

There are multiple ways in which one can explicate a term. One way of separating different attempts is to draw a line between, what I will call, intrinsic and functional characterisations.

Let me start by giving an example. A car can be characterised as an object that allows the transportation of people and goods from A to B. However, one can also characterise a car by giving a description of its number of wheels, engine, fuel efficiency and interior design. The former characterises a car in terms of the functions it performs. The latter defines a car in terms of its intrinsic properties.[14] A parallel distinction can be drawn between inferentialist and substantial accounts of representation. Inferentialist accounts of scientific representation state, roughly, that a model represents a target scientifically if the model allows accurate inferences about the target system (Kuorikoski and Lehtinen 2009). Substantial accounts, as for example Giere's similarity account, try to identify the conditions that need to be in place such that a model can perform its representational function.

[14]The relevant contrast is not between intrinsic and relational properties. I do not intend to make a metaphysical claim about the nature of properties of cars. Intrinsic properties of a car are in an important sense independent from the functions of a car, despite the fact that the car can perform its functions only given some configuration of intrinsic properties.

Accordingly, one can differentiate between a functional and intrinsic characterisation of the notion 'toy model'. A purely functional characterisation would, for example, describe toy models as models from which one can easily gain insights or which are used in particular types of explanations. In contrast, I shall give an intrinsic characterisation of toy models. I find a purely functional characterisation of toy models wanting because it does not illuminate the building blocks and representational relations of toy models. Consequently, functional characterisations are not likely to provide insights that can be used to answer the question that, I think, one should be ultimately interested in: In virtue of what do toy models perform their epistemic roles in different scientific and policy-making contexts?

Let me now introduce my three conditions, which, taken together, should provide a satisfactory intrinsic characterisation of toy models.

### 4.4.2 Manipulability condition

Models share the property that they can be manipulated. One can physically intervene in a material model, such as changing the water inflow in a hydraulic scale model of the San Francisco Bay. One can also manipulate mathematical or graphic models by changing the values of variables or shifting curves, respectively. As Morrison and Morgan (1999, p. 12, 32) point out, the ability to manipulate models is essential for the epistemic import of modelling.

A closer look at my five cases reveals that all these models are manipulable. However, a further peculiarity is striking. These models are manipulable in a *relatively easy way*. Deriving Hotelling's principle of minimal differentiation requires solving a single profit maximisation problem for two producers. The same is the case for Akerlof's car market in which trading breaks down as a consequence of the utility maximising behaviour of buyers and sellers. Schelling's checkerboard model does not require solving any equations but can be simulated with a physical checkerboard and two types of objects by executing the movement rules of the model.

This ease of manipulability is conditional on training in a particular field. In the context of the debate about scientific understanding, de Regt and

Dieks (2005, p. 151) put forward an intelligibility condition that brings out this point. Gaining understanding with the help of a theory $T$ requires that the theory is intelligible. According to de Regt and Dieks (2005, p. 151), a theory $T$[15] is intelligible for scientists in a particular context if "they can recognize qualitatively characteristic consequences of $T$ without performing exact calculations". For example, one can understand the temperature change of a gas confined in a container given a reduction of the volume of the container with the help of the kinetic theory of gases. This is the case since one can recognise the qualitative consequences of this theory. For example, "if one adds heat to a gas in a container of constant volume, the average kinetic energy of the moving molecules (...) will increase." (de Regt and Dieks 2005, p. 152). Importantly, the capacities, background knowledge, and background beliefs of scientists play a crucial role for intelligibility (de Regt and Dieks 2005, p. 151).

These reflections can be condensed into the following manipulability condition:

**Manipulability condition** The handling of a model belonging to a discipline $D$, including the derivation of key model results, is relatively easy for people trained in $D$.

Let me add three clarifications. First, this way of spelling out the manipulability condition captures, in my view, the *simplicity* of toy models, which was particularly highlighted by Hartmann et al. (2016). It is simplicity in terms of manipulability that really sets this model class apart. The fact that the ease of manipulability is relative to the training received in a particular field can be nicely squared with the fact that the structures of toy models can vary quite significantly across different disciplines.

Second, non-toy models, such as a complex dynamic stochastic general equilibrium model or a coupled global climate model, can also be easy to manipulate for the expert in the field equipped with enough training. This is why the condition is formulated in a comparative way: toy models are

[15]de Regt and Dieks (2005) use a broad notion of theory that includes models in the sense discussed here.

*relatively* easy to manipulate. This should be read as relatively easy *compared to* other modelling exercises. So even if manipulating a complex DSGE model is fairly easy for a macroeconomist after enough training, manipulating an economic toy model such as Hotelling's or Akerlof's model is considerably easier.

Third, this way of spelling out the manipulability condition is straightforwardly *agent-centric*. I take this to be a strength of my explication, since if everyone were a Laplacian demon, equipped with unlimited memory storage and computational capacities, modelling projects would look different. The fact that toy models play a prominent role in many sciences must be accounted for in light of our cognitive limitations.

### 4.4.3 Multiple realisability condition

Some results derived with toy models apply to a wide range of target systems. I will capture this observation in a multiple realisability condition. I proceed in three steps to explain this condition. First, I introduce Cartwright's 'ladder of abstraction', which makes talk of different levels of abstraction more precise and allows me to formulate the condition. Second, I show that the condition is satisfied by my five cases. Third, I clarify the condition further by discussing its relation to the generality of toy model results.

Cartwright (1999) discusses how physical laws apply to the world.[16] She starts by pointing out that one can describe the world at different levels. Some of these levels of description are more concrete than others. She then suggests a parallel between the relationship of a fable and its moral to the relationship between a concrete physical situation and a scientific law (Cartwright 1999, p. 36-37). The moral of a fable is an abstract claim (e.g., "the weaker are always prey to the stronger", Cartwright (1999, p. 37)) whereas the fable describes a concrete situation in which the abstract claim is instantiated in a particular way: For example, a marten eats a grouse and gets, in turn, eaten by a fox (Cartwright 1999, p. 39). Accordingly, a scientific law (e.g., Newton's second

[16]Cartwright makes similar points in her later paper *Models: Parables vs. Fables* (Cartwright 2010).

law $F = ma$) is an abstract claim. A particular physical configuration (e.g., a block of wood on a plane that is pushed) is a concrete situation in which the abstract claim is instantiated in a particular way (Cartwright 1999, p. 45).

Cartwright explains this parallel by pointing out that the relationship of abstract to concrete levels of description is straightforward. The concrete description fills out the abstract description (Cartwright 1999, p. 40) as in the case of commonplace language use. For example, saying that "I have worked this morning" is a more abstract description of my cleaning my flat, writing up my thesis, and preparing to teach a class.

This idea of a 'ladder of abstraction' can be used to give a more fine grained account of statements that can be derived from toy models. The results of toy models can be described in concrete terms (e.g., a 9x9 checkerboard showing a particular distribution of black and white dots) or in more abstract terms (e.g., a pattern of segregation). In the following discussion, I refer to properties and relationships described in abstract terms as abstract properties. With this clarification in place, I can formulate the following multiple realisability condition:

**Multiple realisability condition** The abstract properties (and their relations to other abstract properties), which are derived with the model, are instantiated in different ways in target systems.

Importantly, toy models are characterised by this condition in the following sense: Toy models come equipped with the claim that their abstract properties are multiply realised in target systems. This claim can be false. Hence, the multiple realisability condition should not be read as inscribing a success condition onto my characterisation of a toy model.[17]

The five case studies discussed satisfy the multiple realisability condition. The key model result of the DY model is an exponential income distribution. As discussed, many income distributions in the real world show an exponential

[17]The toy model itself instantiates the derived model results in a particular way. The toy model is constructed such that the corresponding behaviour in target systems is illustrated. This puts a constraint on the set of admissible toy models. For example, the Kac ring and not a system in which an agent throws marbles at a wall is chosen, since the former and not the latter exhibits the relevant abstract properties.

bulk. The abstract property of an exponential income distribution is instantiated in different ways in these economies. For example, different currencies are used to save money and different retail banking structures are in places. The claim of the DY model would be false if real world income data did not show an exponential bulk.

The key model result of the Kac ring is the claim that any system with an ergodic decomposition shows an equilibrium-like behaviour. This relation between ergodic decomposability and equilibrium-like behaviour holds for any system with a measure-preserving dynamic. The relation is instantiated in different ways in these measure-preserving systems, such as gases or fluids. The claim of the Kac ring can be false. This is the case for a measure-preserving system that has an ergodic decomposition and does not show an equilibrium-like behaviour.

The key model result of Schelling's checkerboard model is that mild discriminatory preferences can lead to segregation. This relation between mild discriminatory preferences and segregation is instantiated differently in different target systems: segregation patterns in cities with different geographical and economic structures or segregation patterns in different schools or universities. The claim of Schelling's checkerboard model can be false. This claim does not hold in case of social systems in which mild discriminatory preferences are present but in which there is no segregation pattern.

The key model result of Akerlof's market for lemons is that informational asymmetries between buyers and sellers in a market can lead to sub-optimal market outcomes. This relation between informational asymmetries and market outcomes is instantiated differently in different real-world market situations: consumer goods such as cars, medical services, or money lending. The claim of Akerlof's model can be false. This is the case in a situation with informational asymmetries but a working market mechanism.

The key model result of Hotelling's model is that competing agents offer products of all sorts that are very similar. This abstract property of similar offerings from competing agents is instantiated in different ways in different target systems: vendors of different goods such as ice cream or electronic goods, programmes of competing political parties, or religious groups. The

claim of Hotelling's model can be false. This is the case in a situation where one has competing agents that do not supply products that exhibit a high degree of similarity.

Let me clarify the condition further by making explicit its relationship to the generality of a toy model. The target systems to which the toy models are applied vary considerably across the different cases. Take again the Kac ring and the DY model. Although economies and societies in which income distributions can be observed vary considerably (compare, for example, Germany with the US), this variability is smaller than the variability that can be observed across the different systems for which the abstract properties of the Kac ring are true. This variety in the target systems corresponds to the degree of generality of the toy model under consideration. This is a desirable feature of toy models. The generality of a model is not only the desirable feature of explanations but, in the form of 'scope', it is a desideratum of models, hypotheses, and theories more generally (as will be discussed in more detail in Chapter 6 in the context of theory choice). The generality of a toy model is not ex ante fixed. It might be the case that, after further investigations of a target system, it emerges that a system does not exhibit the purported abstract property or, in a different scenario, more target systems exhibit the abstract property than initially assumed.

### 4.4.4 Hybrid representation condition

The manipulability and multiple realisability condition already go some way towards providing a satisfactory characterisation of toy models. However, the two conditions miss one aspect which popped up across the five case studies. In a sense, these models seem to give an obviously *completely wrong-headed picture* of the target systems they are supposed to describe. No one believes that cities are checkerboards on a plane inhabited by black and white dots. It seems far-fetched to claim that Schelling made a mistake by describing cities as checkerboards. He did not need to investigate his model to see that it gives a totally wrong-headed picture of reality. Equally, no one believes that gases are rings with sites occupied by balls that switch their colour upon passing a

marker. Making these intuitions more precise is the task of this section.

The discussion in the previous section has made clear that the question about the representational relationship[18] of toy models and targets needs to be answered for different levels of descriptions. The multiple realisability condition has shown that abstract model results can be instantiated across a variety of target systems. With respect to this level of description, there is a representational relation between toy model results and target systems. However, there is a lack of such a representation relation with respect to the basic entities and properties of toy models. On this level, the toy models portray target systems as checkerboards, rings, or point collisions. In a slightly more precise way, these observations are captured in the following condition:

**Hybrid representation condition** The basic entities, properties, and relations between the basic entities in the model do *not* represent features in any target system. However, there is a representational relation between model results, formulated on a more abstract level of description, and target systems.

This condition is formulated without reference to a particular account of representation. In fact, the condition can be fleshed out with the help of different accounts of scientific representation. I shall discuss two prominent accounts to make the condition more precise.[19] To keep the exposition reasonably short, I will spell out solely one aspect of the hybrid representation condition, that is, the *lack of representation* between the basic level of toy models and target systems.

[18]The notion of representation is ambiguous. One can distinguish between three senses of representation (I follow Contessa 2007, p. 52 here). A model *denotes* a target. Both the logo of the London School of Economics (LSE) and a campus map of the LSE denote the LSE. Denotation can be a matter of pure convention. A model can be an *epistemic representation* of a target in the sense that the model allows performing surrogative reasoning about the target. The map of the LSE campus is an epistemic representation of the LSE campus, whereas the LSE logo is not. Finally, a model can be a *faithful epistemic representation* when it accurately represents the target. With regards to some features, the campus map of the LSE is a faithful representation of the LSE campus. For the subsequent discussion, when I use the term 'representation', I understand it in the sense of epistemic representation.

[19]For a recent and comprehensive review of the literature on scientific representation see Frigg and Nguyen (2017).

### The hybrid representation condition according to Giere's similarity account

Giere proposes a similarity account of representation. The basic idea behind similarity accounts of representation is that scientific models represent target systems in virtue of being similar to them (Frigg and Nguyen 2017, p. 57). According to Giere's account, models come equipped with theoretical hypotheses that assert that a specific model is similar in relevant respects and to specific degrees to a target system (Giere 2004, p. 747). Importantly, these theoretical hypotheses are formed by agents with specific aims and intentions and, hence, an agent-based version of representation results. Giere (2010, p. 274) states this intentional conception of scientific representation as follows:

> Agents (1) intend: (2) to use model, $M$; (3) to represent a part of the world, $W$; (4) for some purpose, $P$. So agents specify which similarities are intended, and for what purpose.

The intentional component is needed to overcome two pressing problems for similarity accounts. First, any object is similar to any other with respect to some criteria of comparison, but it is not the case that anything can represent anything else. Second, similarity is a symmetrical relation whereas representation is a directed relation, that is, models represent targets and not vice versa (Giere 2010, p. 274). By building intentions into the picture both problems vanish: agents specify what similarities are the relevant ones and, by using a model to represent a target, they give the representation relation a direction.

Frigg and Nguyen (2017, p. 60) provide the following helpful formulation of Giere's account:

> A scientific model $M$ represents a target system $T$ iff there is an agent $A$ who uses $M$ to represent a target system $T$ by proposing a theoretical hypothesis $H$ specifying a similarity (in certain respects and to certain degrees) between $M$ and $T$ for purpose $P$.

The hybrid representation condition can be made more precise with the help of the theoretical hypothesis $H$. In the case of a toy model, the claim that the basic entities, properties and relations between the basic entities of the model are similar to the target system is not part of the theoretical hypothesis $H$. To put it more directly, the scientist does *not* claim that the level of basic entities and properties (such as the ring and balls of the Kac ring) stands in a similarity relation to the target system. In contrast, the theoretical hypothesis $H$ contains the claim that the abstract properties of a toy model (such as the entropy profile of the Kac ring) is similar to the entropy profile of a target system (such as a gas or a fluid).

### The hybrid representation condition according to the DEKI account

Frigg and Nguyen (2016) offer an account of scientific representation that emphasises the importance of the interpretation of a model. According to their DEKI[20] account, in a nutshell, a model represents a target system if and only if i) the model denotes the target system, ii) the model properties are endowed with an interpretation in terms of the target system, iii) some of these interpreted properties are exemplified by the model, and iv) these exemplified properties are keyed up and imputed onto the target system. As an example, they discuss the Phillips-Newlyn machine, which is a machine that consists of valves and pipes enabling a flow of water. This machine represents a particular economy, let us say the Swiss economy, if and only if i) the machine denotes the Swiss economy, ii) the properties of the model (such as 'the flow of water') is interpreted in terms of the target system (i.e., 'the money circulation in Switzerland'), iii) some of these properties are exemplified (e.g., a particular level of water in a tank corresponding to an amount of money), and iv) these exemplified properties are turned into final statements about the target system (e.g., the amount of savings in the Swiss economy is such and such) via a key that allows one to add some modification to the exemplified model result (Frigg and Nguyen 2016, p. 19).

The hybrid representation condition can be made more precise in relation

[20]The acronym stands for the different components of their account: Denotation, Exemplification, Keying-up, and Interpretation.

to the operations of interpretation, exemplification, keying up and imputation. The Kac ring can be regarded as a gas representation. For example, the balls of the ring are interpreted as molecules of the gas and the active sites as points in the space where molecules collide. These properties are also exemplified by the model since, roughly, these properties are instantiated by the model and they are epistemically accessible (Frigg and Nguyen 2016, p. 7). However, in an application of the Kac ring, these exemplified properties are not keyed up with properties of the target system. Scientists simply do not care about the individual gas molecules or their collision points when they assess the thermodynamic behaviour of a gas. However, scientists key up and impute the exemplified property of the entropy profile onto target systems of interest.

The discussion of the DEKI account allows me to add a clarificatory remark regarding the context sensitivity of my account of toy models. One might worry that it is in principle possible to impute any exemplified properties onto the target system. Accordingly, it might just be a historical contingency that scientists have not yet keyed up these exemplified properties with a target system. Or, furthermore, it might well be the case that this (lack of) keying up is a context-sensitive matter. With respect to some questions, some of the exemplified properties on the basic level are imputed, with respect to other questions, different ones will be imputed.

I am happy to accept this kind of context-sensitivity of my characterisation of the hybrid representation condition. In fact, the same model can in principle be 'toy' if used by one scientific discipline and 'non-toy' if used by another scientific discipline, if the latter discipline treats the level of basic entities and properties as representationally relevant.

### 4.4.5 Taking stock: A new explication of the term toy model

The discussion in the previous sections allows formulating the following explication of the notion 'toy model'. A toy model is a model type for which the following three conditions jointly hold:

**Manipulability condition** The handling of the model belonging to a disci-

pline $D$, including the derivation of key model results, is relatively easy for people trained in $D$.

**Multiple realisability condition** The abstract properties (and their relations to other abstract properties), which are derived with the model, are instantiated in different ways in target systems.

**Hybrid representation condition** The basic entities, properties, and relations between the basic entities in the model do *not* represent features in any target system. However, there is a representational relation between model results, formulated on a more abstract level of description, and target systems.

Let me elaborate on this characterisation. To start, the multiple realisability condition expresses a new stability criterion: *predictive stability*. As I have argued above, toy models come equipped with the claim that their key model results are instantiated across a variety of target systems. This stability claim can be expressed as an inference scheme as follows: Let $i$ be an index defined over a set of $n$ target systems ($i = 1, ..., k, ..., n$) that differ substantially in their composition and $P(k)$ be the statement that the model result $P$ holds for target system $k$, then predictive stability expresses the following inference:

$$\frac{M \Rightarrow P}{\forall i : P(i)}$$

Predictive stability differs in two respects from derivational robustness analysis. First, derivational robustness analysis is concerned with assessing the implications of alternative *model assumptions* for *model results*. As became clear in the introduction of this thesis, additional conditions need to be in place such that this study of the model world reveals something about target systems. In contrast, robustness across target systems is a claim about the relation of model results and a set of target systems, that is, the results being instantiated across target systems. Second, analysing the robustness across target systems does not require changing the model assumptions. Rather, the focus is on seeing whether a *particular* model exhibits a behaviour that can

be observed across different target systems. This focus follows naturally from the fact that in the discussed cases of robustness across target systems (e.g., for the Kac ring), the basic level of the model (e.g., the ring structure) does not refer to features in the target systems.

The inference pattern of predictive stability shows some similarities with a classical claim from confirmation theory, that is, the claim that a theory or hypothesis is better confirmed by varied evidence (Bovens and Hartmann 2003, p. 103). Consider the following example: The hypothesis 'All ravens are black' is better confirmed by the observation of a black raven in France, Australia, and North America than by the observation of three black ravens in a small french town (Strevens 2017, p. 73-74). Diverse evidence, so one intuition goes, is more powerful in ruling out competing hypotheses, and, hence, has more confirmatory power. Predictive stability does, however, not make a claim about a relation of confirmation. The relationship between the model result $P$ and the instantiation of this result across different target systems is *postulated* by predictive stability. The question whether the model result $P$ is in fact instantiated across different target systems and whether this constitutes confirmatory evidence for $P$ is a *subordinated* question. This will become clearer when I introduce my account of learning with toy models in the next chapter.

My characterisation of toy models also allows addressing the question as to whether there is a sharp distinction between toy and non-toy models. To answer this question, I consider which of the conditions are binary in nature (satisfied/not satisfied) and which can be fulfilled to a lesser or higher degree.

The manipulability condition can be satisfied to a lesser or higher degree, even if one takes disciplinary training into account. Two economic models, both satisfying the multiple realisability and hybrid representation condition, can still differ in their ease of use. I do think that in such a scenario it is intuitive to view one of the two models as more 'toy' than the other. The same line of reasoning does not apply to the multiple realisability condition. If two models are equally easy to manipulate and satisfy the hybrid representation condition, but differ in terms of the number of systems in which the respective abstract properties are instantiated, one of the two toy models will

be more general than the other. What about the hybrid representation condition? Here, some qualification is in order. I want to avoid overgeneralising since I have only looked at five cases of toy models and fleshed out the hybrid representation condition in two ways. However, despite these limitations, my analysis shows that the hybrid representation condition is binary in nature. With respect to all five cases, the level of basic entities and properties of the model is not treated as representationally relevant by the users of the toy model. To sum up, the claims about the non-binary nature of the manipulability condition supports viewing the distinction between toy models and non-toy models not as a sharp one.

Finally, my explication scores well on Carnap's criteria for an adequate explication. First, the proposed explicatum applies to those cases that were identified as clear instances of the explicandum in Section 4.2. Second, the explicatum represents an improvement in exactness as it states three explicit conditions for a toy model whereas each of the conditions is made more precise with the help of established terms (see, for example, the spelling out of the hybrid representation condition in Section 4.4.4). The fruitfulness of this explication will be established in two steps. To start, in Section 4.5, I highlight three corollaries of the explication. In Chapter 5, I show that the explication illuminates the question of the epistemic value of toy models.

## 4.5 Corollaries of the explication

The discussion so far has concentrated on the task of identifying and spelling out conditions to characterise toy models. This conceptual work will pay off in the next chapter in which I put the conditions to use. However, before turning to this task, my explication implies three corollaries that, in my view, are interesting in their own right.

### 4.5.1 Corollary 1: Toy models vs. one-factor models

Given my account of toy models, I can now return to the distinction between toy models and one-factor models that I hinted at when I introduced my case

studies. I argue below that the Ising model, the Lotka-Volterra model, the MIT-Bag model, and Bohr's model of the atom are better described as one-factor models since they do not satisfy the hybrid representation condition. To be more precise, they do not satisfy the first conjunct of the hybrid representation condition, that is, their basic entities and properties do stand in a representational relation to target systems. These one-factor models are, hence, in a crucial respect distinct from the five toy models examined here.

The Ising model, the Lotka-Volterra model, and the MIT-Bag model are regarded as paradigmatic examples of toy models by Hartmann et al. (2016). The upshot of my distinction between toy models and one-factor models is that Hartmann et al. (2016) cast the net too wide with their account of toy models. If these three models are toy models, it is not clear how they can draw a distinction to non-toy models since many models describe solely one (or a few) causal factors and are highly idealised. As I pointed out earlier, Hartmann et al. (2016) do not want to put forward a sharp boundary between toy and non-toy models and regard this classification as a matter of degree. However, my point stands, since even if the distinction is a matter of degree, their notion of toy models can be too inclusive to be a useful category. It is not a useful category since it lumps together too many different types of models and, hence, glosses over important differences between them.

### Ising model

The Ising model (Ising 1925) is a statistical mechanical model of ferromagnetism. Ferromagnetic materials, such as iron or cobalt, can be magnetised by an external magnetic field and remain magnetised after the external source is removed (Cipra 1987, p. 937). In contrast, paramagnetic behavior of a material occurs if the magnetisation is lost progressively as the external magnetic field is removed (Friedli and Velenik 2016, p. 18). The two-dimensional Ising model allows deriving a phase transition from paramagnetic to ferromagnetic behaviour (Friedli and Velenik 2016, p. 28, 30).

In a nutshell, the model works as follows. The material is represented as a two-dimensional lattice. At each point on the lattice there is an atom endowed

with a spin which that is either orientated up or down. The spins interact with each other. According to the Ising model, spins interact only pairwise with their nearest neighbours: the one in the north, south, west, and east (Friedli and Velenik 2016, p. 24). Given these spin interactions, one can define the magnetisation density for the lattice. With respect to this magnetisation density, two observations can be made. First, if the temperature of the lattice system is very high, the magnetisation density is close to zero. This means that the fraction between the up and down spins is essentially equal. If the temperature is very low, the magnetisation density is close to one of the two ground states of the lattice system, that is, a magnetisation of +1 or -1. Hence, in the limit case of very low temperature, one type of spin is favoured and, accordingly, a global order or spontaneous magnetisation is observed (Friedli and Velenik 2016, p. 28). Second, there exists a critical, finite temperature (the Curie temperature) at which this phase transition occurs (Friedli and Velenik 2016, p. 30).

Why is the Ising model not a toy model? The hybrid representation condition is not satisfied. The model captures one of the relevant causal factors underlying the phenomenon of phase-transitions in materials, that is, the spin-alignment (see Weisberg 2007, p. 642-643 and Rice 2016, p. 91). This spin alignment process is based on a quantum mechanic assumption that the magnetic moments in a material (induced by the spin of elementary particles) can be quantized into a two-state system (Knuuttila and Loettgers 2016, p. 383). Due to this, the basic entities and properties of the Ising model (i.e., the spin structure and their alignment) represent features of the target systems of interest. The Ising model even represents these features faithfully. Accordingly, it is *not* the case that solely the pattern of the phase transition of the Ising model stands in a representational relation to the target system.

### Lotka-Volterra model

As stated earlier, the Lotka-Volterra model (Lotka 1925; Volterra 1926) consists of two coupled differential equations that relate the growth of prey and predator populations. The guiding idea is that these quantities are coupled

via a negative feedback: an increase in the number of predators leads to fewer prey and abundance of prey is correlated with the number of predators. The two differential equations can be specified, in the simplest version of the model, as follows:

$$\frac{\partial V}{\partial t} = rV - (aV)P \tag{4.18}$$

$$\frac{\partial P}{\partial t} = b(aV)P - mP \tag{4.19}$$

$V$ and $P$ stand for the size of the prey and predator population, respectively. The change in the prey population over time is expressed as the difference between the prey growth (where $r$ is a constant growth rate) and the prey-capture rate (where $a$ is the prey capture rate per predator). The change of the predator population is expressed as the difference between predator births (which is the prey capture rate per predator multiplied by a constant $b$) and predator deaths (where $m$ is the constant death rate) (Weisberg 2006b, p. 734). The term $aV$ is called the functional response. The term $b(aV)$ is the numerical response.

Why is the Lotka-Volterra model not a toy model? The hybrid representation condition is not satisfied. The Lotka-Volterra model is formulated at the population level. All the parameters in the model (the birth and death rates, the functional and the numerical response) are aggregate terms. This population level is regarded as representationally relevant given one wants to model particular target systems. For example, there are detailed empirical studies which determine the functional and numerical response in predator-prey systems (Holling 1959). Due to this, the basic entities and properties of the Lotka-Volterra model represent features of the target systems of interest. The Lotka-Volterra model, if calibrated correctly, even represents these features faithfully. Accordingly, it is *not* the case that solely the qualitative features of the model (such as the undampenend oscillation) stand in a representational relation to the target system.

### MIT-Bag model

The MIT-Bag model is a model of particles called hadrons. Hadrons, such as neutrons or protons, are composed of quarks. Quantum chromodynamics is the field that studies the strong forces governing the behaviour of quarks (Hartmann 1998, p. 6). Quantum chromodynamics has three features: First, *asymptotic freedom*. Quarks move freely at very high energies, however, their movement is restricted by strong interacting forces at low energy levels. At these low energy levels, one observes, second, *quark confinement*. No single quark is observed because they always come in groups. The third feature is *chiral symmetry*. One does not observe right- and left-handed versions of particle spins, hence, a symmetry breaking occurs in the interaction of these particles, generating the mass of the hadrons (Hartmann 1998, p. 7-8).

The MIT-Bag model is a model of quark confinement. Confinement is modelled as the quarks being forced to move inside a spatial region (the bag) by an external pressure. Within the bag, the quarks occupy particle orbits. No interaction between the quarks is assumed to take place. The simplest version of the model assumes that the bag shape is spherical (Hartmann 1998, p. 9). This model allows recovering some features of hadrons, such as their masses and radii, to some degree of approximation (Hartmann 1998, p. 11).

Why is the MIT-Bag model not a toy model? The hybrid representation condition is not satisfied. The level of basic entities and properties in the model, that is, the quarks in the bag and the bag, is regarded as representationally relevant. The model captures the fact that hadrons consist of two or three quarks. The bag pressure reflects the fact that when a hadron is created, the quarks dig a hole in a non-perturbative vacuum populated by gluons and other entities. The inside pressure of a hadron must be as high as the outside pressure to guarantee the observed stability of hadrons (Hartmann 1995, p. 8). Due to this, the basic entities and properties of the MIT-Bag model represent features of the target systems of interest. The model even represents these features faithfully. Accordingly it is *not* the case that solely the qualitative features of the model (i.e., the features of hadrons, such as their masses and radii) stand in a representational relation to the target system.

### Bohr's model of the atom

As stated earlier, Bohr's model of the atom is a model of particle physics. The model superseded a variety of models of the atom that were proposed at the beginning of the 20th century. For example, Thomson's model describes the atom as a uniformly positively charged sphere inside which the negatively charged electrons move in circular orbits held together by electromagnetic forces (Bohr 1913, p. 1). However, around this time it was discovered that electrodynamics cannot be fruitfully applied to systems as small as atoms. In contrast, Bohr's model makes space for quantum effects and assumes that an atom consists of a nucleus and electrons orbiting the nucleus in concentric trajectories (Bohr 1913, p. 2, 4).

The key insights drawn from Bohr's model is the structure of Rydberg's formula as well as an expression of the Rydberg constant in terms of more fundamental constants of nature (Bokulich 2011, p. 41). Rydberg's formula describes the spectral lines of a hydrogen atom. A spectral line is light of a specific frequency that is emitted from atoms or molecules when a change of energy happens (Bokulich 2011, p. 41).

Bohr's model is now regarded as superseded by the more accurate quantum mechanical model of the atom (i.e., the valence shell model). The key difference is that the valence shell model does not assume that electrons orbit nuclei on trajectories. In fact, the quantum mechanical model of the atom states that there are no trajectories (Bokulich 2011, p. 41).

Roughly, how does Bohr's model work? Electrons are assumed to orbit a nucleus in a discrete series of classical trajectories. These trajectories are called stationary states. If an electron is in a stationary state, the energy of the electron is constant. If the electron moves (or "quantum jumps") from one stationary state to another, the electron loses or gains energy. If such a move occurs, a photon of a given frequency is emitted. The frequency is determined by the energy differences between the orbits involved in the move (Bokulich 2011, p. 41).

Why is Bohr's model of the atom not a toy model? The hybrid representation condition is not satisfied. The level of basic entities and properties in

the model, that is, the nucleus, the electrons, as well as the orbits, is regarded as representationally relevant. Accordingly it is *not* the case that solely the qualitative features of the model (e.g., the Rydberg formula) stand in a representational relation to the target system. The fact that Bohr's model was superseded by a model that does not contain electron orbits reveals that this aspect of the basic structure of Bohr's model does not represent an aspect of the target system faithfully. This is, importantly, not the same as treating these orbits from the outset not as representationally relevant. It took time and hard work from particle physicists to establish whether electrons orbit nuclei or not on these trajectories. When Bohr introduced the model, it was not clear that the orbits gave a totally wrong-headed picture of the atom.

### 4.5.2 Corollary 2: Toy models vs. probing models

My explication allows differentiating between toy models and what I suggest calling probing models.[21] Probing models are, in contrast to toy models, not introduced in relation to any (type of) real world target system. Rather, they are used to explore theories or nomologically impossible model worlds. Let me introduce two examples to further clarify the distinction between a toy and a probing model.

#### phi 4-model

Quantum field theory is a theoretical framework that is used to construct quantum mechanical models of subatomic particles and quasi-particles in condensed-matter physics. According to this framework, quantum mechanical interactions between particles are expressed as interactions between quantum fields (Stefanovich 2014, p. 299). The $\varphi_4$-model is the simplest quantum field theory that can be constructed. It shows a series of interesting features that are also shared by more complex quantum field theories: the $\varphi_4$-model allows

[21]The term 'probing models' is also used by Frigg and Hartmann (2012, p. 11). Note, however, that they use the terms 'toy model' and 'probing model' interchangeably.

introducing the technique of renormalization[22] and is Lorentz invariant[23].[24]

According to Hartmann (1995, p. 9), the $\varphi_4$-model does not represent anything. He suggests that it rather gives physicists a "feeling" for what quantum field theories are like. In my view, this metaphorical talk about "getting a feeling" or "getting a handle of something" is best spelled out in the following way: probing models, such as the $\varphi_4$-model, allow *exploring* the features of theories. With the help of these models it can be asked what properties classes of theories share and these shared properties can be investigated further in the simpler setting of the probing model, as is the case with the Lorentz invariance that is exhibited by the $\varphi_4$-model and more complex quantum field theories.

### Ratchet and pawl machine

A perpetual motion machine is a machine that can produce work indefinitely (Weisberg 2013, p. 126-127). The ratchet and pawl machine is particular model of a perpetual motion machine. Two boxes are connected by an axle. At the one end of the axle, four vanes are attached and contained in the box on the right with temperature $T_1$. At the other end of the axle, a ratchet is attached and contained in the box on the left with temperature $T_2$ (see Figure 4.7).

Due to a dampening pawl, the ratchet can only turn in one direction and releases heat into the box on the left when it turns. In the middle of the axle there is a wheel with a weight attached to it. When the wheel turns and the weight is lifted, the machine is doing work. Now, how could this mechanism work? If the box on the right is filled with a gas, then the gas molecules

[22]Very roughly, renormalization is a mathematical technique that enables dealing with infinities that arise in expressions for measurable terms in quantum mechanics due to the number of underlying particles and their interactions in a quantum system (see Li 2012 for an introduction).

[23]Very roughly, Lorentz invariance of a system of equations denotes the fact that if the equations hold in one inertial reference frame then they hold in any inertial reference frame (see Mattingly 2005 for details).

[24]Contrary to my account, Hartmann et al. (2016, p. 8) regard the $\varphi_4$-model as an embedded toy model. In my view, this is inconsistent with their claim that toy models are target-directed (see Hartmann et al. 2016, p. 3).

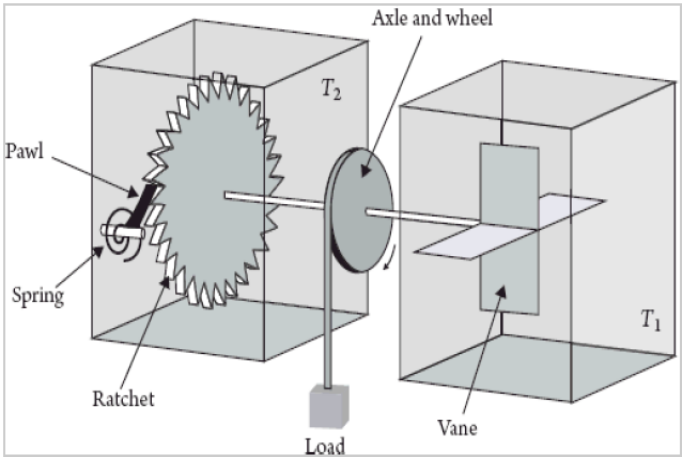


Figure 4.7: Visualisation of the ratchet and pawl machine (Hartmann et al. 2016, Figure 7.3).

can push the vanes in one direction, thereby turning the axle in the direction allowed by the ratchet, and, hence, the weight is raised and work is done. Since there are no perpetual motion machines, the model can be used to show why this is the case. In particular, it allows identifying which laws need to be different such that the machine would work. To see the trouble, consider the assumption that $T_1$ is equal to $T_2$. If the two boxes have equal temperature to start with, then the turning of the ratchet and the associated dampening mechanism rises $T_2$. A rise in $T_2$ means that the molecules in the box on the left are pressing against the pawl and, from the right angle, could lift it and reverse the ratchet wheel, thereby undoing the work (Weisberg 2013, p. 128). The machine could only work if heat is extracted from the box on the left but this is an extra process and one no longer has a perpetual motion machine. Accordingly, the laws of energy conversion had to be different if the machine is supposed to work. A form of violating this law could be the assumption that energy just disappears.

Let me take a step back and reflect on the functioning of this type of probing model. In contrast to the $\varphi_4$-model, the ratchet and pawl machine is not introduced in relation to a particular theory; although thermodynamics is the relevant theory in the background here, the model is not introduced to study thermodynamics. The ratchet and pawl machine postulates a nomological

impossible target. By postulating this scenario, it provokes a fruitful comparison between actual or possible real world targets. It is this comparison which enables the insight about the status of the energy conservation principle. However, this comparison with possible or actual models is not necessary to gain a conceptual insight from a probing model of this sort. Exploring a nomologically impossible model worlds can also yield conceptual insights.

### 4.5.3 Corollary 3: The problem of de-idealising toy models

My explication of the notion 'toy model' provides a handle on the observation that it is often not straightforward to de-idealise toy models. By de-idealisation I refer to the process of turning a model into a more accurate description of a particular target system or type of target system. McMullin (1985) offers a helpful framework for de-idealisation. *Formal de-idealisation* is the process of adding back explanatory factors which were seen as relevant in the first instance of model building but were not included for tractability reasons (McMullin 1985, p. 258). McMullin gives the example of the movement of the sun that is omitted in the first versions of Newton's model of the solar system although Newton was well aware of the impact of the sun's movement on the trajectory of the earth. *Material de-idealisation* is the process of giving additional details about an explanatory factor that is already included in the model (McMullin 1985, p. 258). McMullin discusses the theory of kinetic gases which does not model the internal structure of molecules, although their internal structure can become relevant in the context of a different line of inquiry.

Let me make my initial observation more precise by distinguishing two questions. First, one can ask whether it is difficult to de-idealise a *particular* toy model to arrive at a non-toy model. Second, one can ask whether there is something about the category of toy models that prevents de-idealisation being straightforward. I address the second question and argue that the answer is yes.

My explication reveals two reasons why de-idealising a toy model is not

a straightforward task. To start, it might not be clear which of the target systems (or types of target systems) should be chosen as the target for the de-idealised model. Take the DY model as an example. A de-idealised DY model looks quite different if the target system is Swiss society as compared to US society, with the modelling aim of reducing income inequality. The difference consists in the set of relevant explanatory factors which are introduced in the non-toy version of the model. There are likely to be different institutional factors relevant in the case of Switzerland and the US, for example.

One might object here that this problem is usually settled by the context of investigation. If one decides to de-idealise a model, this does usually happen in the context of particular modelling tasks. For example, one can de-idealise the DY model by taking into account income from wealth and the educational structure in a society to arrive at a model better suited to guide policy interventions to reduce income inequality.

However, even if the context of investigation can settle this objection, there remains one further problem for de-idealising toy models. These models contain a level of non-representing basic entities and properties. It is unclear what the de-idealisation for these basic entities and properties looks like. How should one de-idealise the ring in the Kac ring, or the checkerboard structure in Schelling's model? A salient answer is to look for guiding principles for the de-idealisation. These seem to be available in cases where the toy model is embedded in a theoretical framework.

Note, however, that not all toy models are "embedded" in Hartmann et al. (2016)'s sense (see Cartwright 2009 for raising this point early in the debate). A toy model is embedded if and only if it is a model of an empirically well confirmed framework theory (Hartmann et al. 2016, p. 5). As Hartmann et al. (2016) point out, there are also "autonomous" toy models, that is, models that are not models of a well-confirmed framework theory. My case studies can serve as illustrations here. The Schelling model and the Akerlof model are autonomous toy models. There is no well-confirmed framework theory (either an overreaching sociological framework in the case of Schelling, or a well-established economic theory in the case of Akerlof). Hence, there is no framework that can provide clear guidance on the de-idealisation process.

Furthermore, even if there were a theory that guided the de-idealisation with respect to one type of target systems, it is not clear that this theory could also guide de-idealisation with respect to different types of target systems.

The lack (or presence) of a guiding theoretical framework marks an additional difference between my cases of toy models and the Ising, Lotka-Volterra, and MIT-Bag models. The Ising model is embededded in a theory of non-relativistic quantum mechanics (see Hartmann et al. 2016, p. 8) which allows for a straightforward de-idealisation of the models by taking into account additional factors. The Lotka-Volterra model is widely regarded as the simplest possible model of predator-prey interactions (Briggs and Hoopes 2004, p. 299) whereas more complicated models in mathematical ecology are de-idealisations of this model. Furthermore, Hartmann (1998, p. 11-12) discusses three approaches of a theory guided de-idealisation of the MIT-Bag model by, for example, taking into account the interaction between the quarks within a hadron. The Bohr model is a slightly different case. As Bokulich (2011, p. 43) argues convincingly, the Bohr model cannot be straightforwardly de-idealised. The reason, however, is not the fact that its basic entities and properties do not stand in a representational relation but the fact that a key component of the model (the orbit assumption) turned out to be false. More accurate models of the atom cannot build upon the Bohr model with respect to this assumption and, hence, need to conceptualize the movement of electrons differently. This, as turned out to be the case, required taking into account a quantum-mechanical understanding of electrons.

## 4.6 Concluding remarks

In this chapter, I provided a new characterisation of 'toy models' involving a manipulability, a multiple realisability, and a hybrid representation condition. This characterisation enabled me to address my first research question: Do measurement and derivational robustness analyses exhaust the set of useful types of robustness analysis? I argued that the answer is no and that predictive stability denotes an important additional category of robustness.

Furthermore, I showed that my explication implies three interesting corol-

laries. First, it enables differentiating between toy models and one-factor models that do not satisfy the hybrid representation condition. Second, it allows distinguishing clearly between toy models, which are target-directed, and probing models, which postulate nomological impossible worlds or are used to explore theories. Third, the explication allows giving a sense of why toy models are hard to de-idealise.

The conceptual scaffolding work of this chapter forms the basis for investigating the natural follow-up questions: how and what can one learn based on toy models? I turn to these questions in the next chapter.

# Chapter 5

# Robustness in Economic Policy-Making: Learning Based on Toy Models

## 5.1 Introduction

In the previous chapter, I identified a set of paradigmatic toy models and characterised them with the help of three conditions: a manipulability, multiple realisability, and hybrid representation condition. My analysis revealed that toy models exhibit a particular form of robustness – predictive stability, that is, the claim that key model results are instantiated across a variety of target systems. In this chapter, I explore what kind of learning toy models allow. A particular emphasis will be placed on whether learning provided by these models can be used in economic policy-making. As my discussion will make clear, many toy model results are derivationally non-robust. Hence, I can address my third research question: What should one do if model results are derivationally non-robust?

To address this question, I proceed in a stepwise fashion. In particular, I defend three claims. First, one must distinguish between learning *from* a toy model and learning *with* a toy model. Learning from a toy model is a form of conceptual learning, that is, an insight into what certain assumptions imply.

Learning with a toy model is learning about a target system, but here the toy model is merely a tool in the learning process. The actual learning happens in the comparison of toy model results with a target system. This learning via comparison takes place if the abstract property (or relation between abstract properties) posited by the toy model is instantiated in target systems that are distinct from the initial system that motivated the construction of the toy model. Second, the issue of the derivational robustness of toy model results has important implications for learning based on a toy model. Non-robust toy models still can provide conceptual learning or can introduce a potentially valuable modelling technique. Third, from these two claims it follows that toy models can only play a limited role in economic policy-making: toy models do not reveal precise predictions about or stable relations that can be exploited for interventions in particular target systems.

This chapter is structured as follows: To start, I review existing accounts of model-based learning to identify conditions for successful learning (*Sections 5.2*). I then introduce the distinction between learning from a toy model and learning with a toy model (*Section 5.3*). Then, I put forward my account of learning from and with toy models (*Section 5.4*) and discuss what the value of this type of model-based learning is for economic policy-making (*Section 5.5*).

## 5.2 Existing accounts of model-based learning

To arrive at an account of learning based on toy models, I start by examining three prominent approaches in the literature that propose accounts of model-based learning. My strategy is, for each of the three approaches, to uncover conditions for model-based learning and to state them as precisely as possible. I then use these conditions to motivate and flesh out my account of learning for toy models in subsequent sections.

I will discuss Grüne-Yanoff (2009, 2013a)'s, Hartmann et al. (2016)'s, and Battermann and Rice (2014)'s accounts. I have selected these three accounts since their intended domains of application covers types of models that share

important features with my paradigmatic cases of toy models.[1]

### 5.2.1 Grüne-Yanoff's account of learning with minimal models

Recall Grüne-Yanoff (2009)'s claim that one learns from minimal models because results derived with the help of these models can affect one's degree of belief in impossibility hypotheses. Accordingly, Grüne-Yanoff suggests that model-based learning consists in a change of degree of belief about particular types of hypotheses. Importantly, Grüne-Yanoff views this kind of learning as a learning about target systems (Grüne-Yanoff 2009, p. 81, 85). A derivation with the help of a minimal model can affect one's degree of belief since this derivation displays a relevant possibility for a target system of interest. Grüne-Yanoff's claim is that a derivation *solely* based on a minimal model can ground learning about a target system.

What conditions need to be in place for this type of learning? As has already been hinted at in the previous chapter, Grüne-Yanoff suggests that minimal models need to be *credible* to allow the derivation of relevant possibilities. Credibility is spelled out as a coherence of the model set-up with the intuitions of the epistemic agent who uses the model. According to him, it is explicitly not the case that one needs to look at the coherence of an agent's *model-independent* intuitions with model assumptions. Rather, Grüne-Yanoff puts forward a purely model-internal notion of coherence. The reason for this is his claim that the intuitions of model users often do not exist independently of the model world (Grüne-Yanoff 2009, p. 94).

Let me look closer at this condition for successful learning about target systems. Is a model-internal notion of credibility enough to establish that a model derivation is a relevant possibility for a real-world target? Fumagalli

[1]Note that the two questions about the characterisation of toy models (what is a toy model?) and their epistemic value (what does one learn based on toy models?) are distinct. In Chapter 4, I argued that Grüne-Yanoff's, Hartmann et al.'s and Batterman and Rice's accounts by themselves do not provide a satisfactory answer to how my paradigmatic cases of toy models should be characterised. This does not rule out that their respective accounts of model-based learning can be applied to toy models.

argues that this is not the case. He succinctly notes this for the case of Schelling's checkerboard model (Fumagalli 2015, p. 17-18):[2]

> However, it [the model] does not per se foster justified changes in confidence in hypotheses concerning real-world segregation processes. To be sure, modellers may occasionally be able to demonstrate that the possible cause of abstract segregation identified by Schelling's model can foster segregation also in the real-world situations they investigate (e.g., think of cases where independent studies provide modellers with this information). Still, on the supposition that Schelling's model is minimal, this demonstration would require modellers to supplement such a model with information or presuppositions regarding those real-world situations (...).

Grüne-Yanoff (2013a, p. 853) seems to acknowledge the need for such a model-external step for establishing the claim that a relevant possibility is provided by a minimal model. He makes a very short remark about how such a link can be established. According to Grüne-Yanoff, such a link is present if the model assumptions are conceptually linked to real world features. Grüne-Yanoff explicitly states that this conceptual link does not presuppose a resemblance relation between the model and the real world. He states that such a conceptual link is given, if, for example, a "possible migration rule in an agent-based model falls under the same concept as actual agents' decision rules concerning migration." (Grüne-Yanoff 2013a, p. 853). One way of spelling out these brief remarks is the following: One can describe both the model world and the real world on different levels of abstraction. There is a level of abstraction at which a phenomenon of the real world and a model assumption are instantiations of the same abstract notion. Let me elaborate on Grüne-Yanoff's example in this spirit. Real world agents have reasons for migrating. One of these decision rules could be to move if the neighbourhood does not provide adequate primary schools. Agents in Schelling's checkerboard are endowed with a decision rule for moving between squares (i.e., move if a certain

[2]See Casini (2014, p. 649) for the same claim about Grüne-Yanoff's notion of credibility.

percentage or more of agents in the agent's neighbourhood are of different type). Both these decision rules can be described in more abstract terms. The two rules are, for example, instantiations of the abstract notion of "being dissatisfied with composition of the neighbourhood".

The key question is now, however, what kind of link between a model and a target system is established in this way. According to Grüne-Yanoff, the relation describes the fact that two phenomena can be viewed as instantiations of the same abstract notion. This establishes the fact that situations can be described with similar vocabulary. Is this enough to show that model results are a relevant possibility for the target system? I do not think so. Consider again the example from Schelling's checkerboard model. What one needs to establish is the claim that factors for segregation identified in Schelling's model (i.e., mild discriminatory preferences) can lead to segregation in real world targets. For this, it is not enough to describe factors in the model and aspects of the real world phenomena with the same abstract notion. Even if this re-description can be done, it is still an open question whether the identified aspects of the real world phenomena are sufficient for leading to segregation in real world targets.

Sugden (2000) provides a different account of the credibility of models. Importantly, this account is *not* based on a model-internal notion of credibility. According to Sugden, the credibility of a model is the compatibility of model assumptions with known general laws governing events in the real world (Sugden 2000, p. 25). Accordingly, our model-independent background knowledge comes into play here. Note that what Sugden is stressing here are relevant similarities between a model and a target system. Relevant similarities hold between the dynamics of a model and the laws governing the dynamic of the target system. Hence, Sugden provides a similarity-based account of credibility or relevance of a model for a (type of) target system. These similarities justify, in Sugden's view, the "inductive leap" from the model result to claims about a target system of interest (Sugden 2000, p. 20).

What emerges from this discussion is the insight that there is a gap between model results and claims about target systems. If one wants to learn something with the help of a model about a real world target, one needs to

be able to bridge this gap. One such way is to show that model results are *relevant possibilities*. As my discussion of Grüne-Yanoff showed, such a relevance relation cannot be established in a purely model-internal way. Sugden's suggestion – establishing relevance via similarity between the model dynamic and causal laws governing the target – is a more promising way of bridging this gap.

### 5.2.2 Hartmann et al.'s account of understanding with toy models

Recall Hartmann et al. (2016)'s account of gaining insights via toy models, which is based on the notions of understanding and explanation. They claim that an individual scientist $S$ understands a phenomenon $P$ via model $M$ in context $C$ if one of the following conditions holds:

- $S$ has how-actually understanding of phenomenon $P$ via model $M$ in context $C$ if model $M$ provides a how-actually explanation of $P$ and $S$ grasps $M$.
- $S$ has how-possibly understanding of phenomenon $P$ via model $M$ in context $C$ if model $M$ provides a how-possibly explanation of $P$ and $S$ grasps $M$.

Here, Hartmann et al. reveal that the key notions of their account are how-actually and how-possibly explanations. Hence, they provide an explanation-based account of learning. The distinction between how-actually explanations and how-possibly explanations can be put as follows. How-actually explanations of phenomena consist of the true (or approximately true) explanatory factors of phenomena (Hartmann et al. 2016, p. 18). In contrast, how-possibly explanations involve merely possible explanatory factors that could account for the explanandum phenomena (Hartmann et al. 2016, p. 18). The how-actually explanation of a phenomenon is a subset of the how-possibly expla-

nations of this phenomenon.[3]

What conditions need to be in place for this type of learning? As stated in the previous chapter, Hartmann et al. distinguish between embedded toy models, which can deliver how-actually explanations, and autonomous toy models, which can deliver how-possibly explanations. An embedded toy model only delivers how-actually explanations if the following conditions are met (Hartmann et al. 2016, p. 19-20). First, the embedding framework theory permits an interpretation and justification of the idealisations of the model. Second, this interpretation and justification needs to be compatible with the veridicality condition, which states that explanatory assumptions are required to be true or approximately true (Hartmann et al. 2016, p. 15). As an example, they provide Newton's Sun-plus-one-planet-model. The idealisations of this model are justified for pragmatic reasons since they turn calculating earth's orbit into a mathematically tractable problem. Furthermore, Newtonian mechanics offers a way of de-idealising this model (e.g., by adding the influence of other planets and moons) such that the veridicality condition is satisfied (Hartmann et al. 2016, p. 20-21). Hartmann et al. (2016) are less explicit about the conditions that need to be in place for how-possibly explanations. They say merely that a how-possibly explanation is a potential explanation of a general pattern. Importantly, they state, the models that provide how-possibly explanations cannot themselves inform us as to whether they have identified the actually present explanatory factors.

Hartmann et al. (2016, p. 27) discusses three functions of how-possibly explanations: a modal, a heuristic, and a pedagogical function. For my purposes, the modal and the pedagogical are crucial. The modal function of how-possibly explanations is identical to Grüne-Yanoff's claim about influencing one's beliefs about impossibility claims regarding target systems (Hartmann et al. 2016, p. 27). If my line of reasoning regarding Grüne-Yanoff's account is compelling, then this modal function of how-possibly explanations also requires an additional step to establish that a model displays a relevant

[3]I deliberately brush over the question here as to whether there is for every explanandum exactly one how-actually explanation. For an entry point into this debate see Pincock (forthcoming).

possibility. To put it differently, one needs to be able to give an argument for why a particular how-possibly explanation should be part of the set of relevant how-possibly explanations with respect to which the true how-actually explanation is a member. The pedagogical function of how-possibly explanations is to enable students and researchers to quickly grasp the idea behind the solution to a problem, to describe a phenomenon, or to apply calculation techniques (Hartmann et al. 2016, p. 27). Importantly, (Hartmann et al., 2016, p. 27) state that this function does not presuppose establishing a link between a target system and the model under consideration. This function can be performed by the model itself.

Two insights emerge from this discussion. To start, learning qua providing how-actually explanations or modal insights via how-possible explanations requires establishing a link between a model and a target system. This link cannot be supplied by the model itself. Furthermore, there is model-based learning qua providing pedagogic tools which does not presuppose such a link. This is a non-target-directed type of learning.

### 5.2.3 Batterman and Rice's account of minimal model explanations

Recall Battermann and Rice (2014)'s argument that minimal models can provide minimal model explanations. In contrast to Hartmann et al. (2016)'s account, a different notion of explanation is in play here. In particular, minimal model explanations, they claim, do not explain in virtue of pointing out features that are shared between a model and a target system of interest. Rather, the connecting element between a minimal model and a target system is the fact that they belong to the same universality class.

What conditions need to be in place for this type of learning? As stated in the previous chapter, showing that a model and a target system belong to the same universality class involves answering the following questions (Battermann and Rice 2014, p. 150):

1. Why are the common features among systems necessary for the phenomenon to occur?

2. Why are the remaining heterogeneous details (those left out of or misrepresented by the model) irrelevant for the occurrence of the phenomenon?

3. Why do very different target systems share common features?

Putting it positively, these three questions reveal the conditions that need to be in place for having a minimal model explanation. One needs to be able to tell why heterogeneous details in the target systems belonging to a universality class are irrelevant; why the shared features among these target systems are necessary; and, finally, why their common features are shared across different target systems. Applying the technique of group renormalisation, which was introduced in the previous chapter, is a way of satisfying these conditions.

Importantly, group renormalisation thereby plays the role of establishing a link between the model and the target systems that lie in the explanatory domain. To see this, recall that the model system is just one additional system that enters into group renormalisation besides the real world (and potential) target system(s). As Lange (2014, p. 295) points out, however, even if a toy model and a target system belong to the same universality class, it is not clear that a satisfactory explanation has been given. Battermann and Rice (2014)'s account fails to establish the asymmetry of minimal model explanations: minimal models explain a target phenomenon and not vice versa. Batterman and Rice cannot claim that minimal model explanations lack explanatory asymmetry. For scientific explanations involving minimal models exhibit explanatory asymmetry despite the fact that the asymmetry is not generated by causal factors figuring in the explanations (minimal model explanations, according to Battermann and Rice (2014), do not explain in virtue of the fact that they accurately represent (relevant) causal factors) (Lange 2014, p. 296-298). One possible way out for Batterman and Rice could be to turn their account into an explicitly agent-based account of scientific explanation. In such a version of the account, Batterman and Rice could stipulate that an agent intends to show that a model and a target system belong to the same universality class and that she uses the model to represent the target system. By this the agent introduces with her behaviour the relevant explanatory asymmetry.[4] I think

[4]See the discussion of Giere's similarity account in Section 5.4.4 for the same move.

that such a move is a compelling way of defending the account against Lange's objection because I believe one should reserve an important role for the epistemic context and the epistemic agents operating within, as will become clear in Chapter 6 on model and theory choice.

Regardless of whether one thinks Battermann and Rice (2014)'s account can be saved from Lange's objection by this move, one important insight emerges from this discussion: Even a form of model-based explanation which locates the explanatory power of models not in shared features between models and targets, comes equipped with a way of establishing a link between a model and a set of target systems. This supports the points made in relation to Grüne-Yanoff's and Hartman et al.'s accounts: Learning something about a target system with the help of a model requires a way of establishing the relevance of the model results for a particular target. This justificatory step cannot be skipped.

## 5.3 Taking stock: Some clarifications regarding model-based learning

So far, I have used the term "model-based learning" as an umbrella term that covers different forms of model-based explanations and the adjustment of degrees of beliefs in impossibility claims. In this section, I rely on the insights from the discussion so far to arrive at a more specific notion of model-based learning.

Two key insights emerge from the discussion so far. First, one should distinguish between target-directed and non-target-directed accounts of learning. Target-directed accounts view models as supplying statements about (types of) real world targets. Non-target-directed accounts see models as yielding statements that should not be viewed as claims about target systems. The two types of learning come with different justificatory requirements. Target-directed learning requires establishing a link between model statements and statements about target systems. As the discussion in the previous section made clear, this link needs to be established with the help of facts that are not

to be found in the model-world itself. Second, not all forms of model-based learning involve explanations. Grüne-Yanoff's account solely involves an agent changing her degree of beliefs and the pedagogical function of how-possibly explanations also does not involve a reference to explanations.

Taking these two insights on board allows me to introduce the key distinction of this chapter: the distinction between learning from a toy model and learning with a toy model.

*Learning from a toy model* is present if the toy model itself provides the learning. The insights that are gained in this type of learning do not require additional input, that is, the facts that can establish a link between a model and a target system. The previous discussion reveals that the pedagogical function of how-possibly explanations are an instance of learning from a toy model. *Learning with a toy model* is present if the toy model itself does not provide the learning but nevertheless plays an important role in the learning process. The toy model is a tool in this learning process. The discussion in Section 5.2 shows that target-directed learning is an instance of learning with a toy model.[5]

I take target-directed model-based learning to consist in the following: A justified change of an epistemic agent's degree of belief in one or more hypotheses about a (type of) target system that makes essential reference to a model. Let me clarify this characterisation.

First, this notion of learning presupposes a justification of a change in belief. This justification can take various forms across different types of target-

[5]Let me further clarify how the two distinctions of target-directed vs. non-target-directed and learning from a model vs. learning with a model relate to each other. Taking the discussion in Section 5.2 at face value reveals that the combination of learning from a model/target-directed learning is ruled out. However, the combination learning with a model/non-target-directed is possible. This would be the case if a model plus some additional facts, for example, a de-idealisation, revealed some additional conceptual insights. The two main categories are, however, "learning from a model/non-target-directed" and "learning with a model/target-directed". This distinction is different from Sugden's distinction between learning about a model world and learning about a target system (see Sugden 2000). Although "learning from a model/non-target-directed" does not involve a target system, the upshot of this kind of learning does not consists in understanding a particular model description (or "model world") but rather in grasping a relation between concepts. To see the difference, note that these conceptual relations could be instantiated in other model descriptions.

directed model-based learning.[6] Second, let me emphasise that this notion of learning allows explicitly for non-explanatory learning, that is, one can learn something about a (type of) target system without being in possession of an explanation.[7] Third, this notion of learning is compatible with accounts of representation that stress the fact that model results are usually not directly applied to target systems. The DEKI account (see Section 5.4.4) is such an account. According to this, a key is used to translate model results into statements that are then imposed onto a target system. Hence, the hypothesis that I refer to in my definition does not need to be identical to the derived model result.

## 5.4 A new account of learning from and with toy models

In this section, I spell out my account of learning based on toy models. After introducing the key tenets of the account (Sections 5.4.1-5.4.2), I strengthen it by articulating two of its features. First, the account allows stating how there can be learning despite the fact that the basic entities and properties of toy models do not represent features of target systems (Section 5.4.3). Second,

[6] If model-based learning involves providing an explanation of a phenomenon, different types of explanations require different justifications for the change in the degree of belief. For example, causal explanations require, very roughly, that one has identified key causal factors contributing to the phenomenon. In contrast, unificationist explanations require, very roughly, that one shows how the phenomenon fits into a pattern involving a range of different phenomena.

[7] After all this conceptual work, the reader might be puzzled why a straightforward question has not been asked yet: Are any of the discussed accounts of model-based learning capable of capturing the epistemic import of my five paradigmatic examples of toy models? I have suppressed this question so far to have all the components in place to give a succinct answer. As accounts of learning with a toy model about a target system, the accounts of Grüne-Yanoff (2009), Hartmann et al. (2016), and Battermann and Rice (2014) can only be applied to my cases if some additional, model-external facts establish a link between the models and the intended target systems. For example, solely if the group renormalisation technique were to work for the Kac ring, then the Kac ring could provide a minimal model explanation of some universally observable behaviour across target systems. Similarly, solely if one can provide a reason for why Schelling's checkerboard model yields a relevant possibility, this model can be said to provide a how possibly explanation of a target phenomenon.

the account enables saying precisely what role the robustness of toy model results plays for learning with these models (Section 5.4.4).

### 5.4.1 Learning from a toy model

One learns from a toy model if the toy model itself provides the learning. In particular, toy models enable one to learn that certain configurations of assumptions have particular implications. Uncovering these inferential relations amounts to one grasping a conceptual scheme.[8] Let me illustrate these rather abstract claims with the help of a few of my case studies.

The key model result of the Kac ring is that a system with an ergodic decomposition shows equilibrium-like behaviour. The Kac ring highlights the concepts of thermodynamic equilibrium, ergodic decomposability, deterministic mechanics and time-reversibility. The model itself reveals relations between these four abstract notions. In particular, the model shows that the assumption of ergodic decomposability, deterministic mechanics, and time reversibility implies an equilibrium like behaviour. So, the Kac ring in itself provides an insight into the inferential relations between these four concepts, and, due to its ease of manipulability, does this in a transparent way.

The key model result of the DY model consists in an exponential income distribution. Looking at the DY model makes clear that the model highlights the concepts of random selection, resource exchange rules, entropy maximisation, and resource distribution properties. The model itself reveals relations between these four abstract notions. In particular, the model shows that the assumptions of random selection, pooling and random division of resources, and entropy maximisation imply an exponential distribution. So, the DY model itself provides an insight into the inferential relations between concepts, and, due to its ease of manipulability, does this in a transparent way. In fact, the model was used as a starting point for exploring further conceptual relations. Adding an assumption about resource saving decisions reveals

[8]Note that I use the notion of a conceptual scheme in a less demanding sense than Davidson (see Davidson 1973). Davidson denotes something akin to a world view with this notion whereas I have a set of concepts in mind that stand in inferential relations to each other.

that the resulting distribution is not only exponentially distributed but also has a power law tail.

The key model result of Akerlof's market for lemons is that informational asymmetries between buyers and sellers in a market can lead to sub-optimal market outcomes. Looking at Akerlof's market for lemons makes clear that the model highlights the concepts of the information states of market participants, trading schemes, and the optimality of market outcomes. The model itself reveals relations between these three abstract notions. In particular, the model shows that the assumptions of informational asymmetries between two market sides and a willingness-to-pay based trading scheme result in a lack of market interactions. So, Akerlof's market for lemons itself provides an insight into the inferential relations between concepts, and, due to its ease of manipulability, does this in a transparent way.

Two things are important with respect to these examples. First, these inferential relations are relations between concepts and should not be read as statements about (types of) target systems. This form of learning with toy models is non-target-directed. Second, this does not preclude that these conceptual resources are used to arrive at explicitly target-directed models that are non-toy in nature.

### 5.4.2 Learning with a toy model

Policy-making involves real world problems and, hence, a pressing question is whether toy models allow learning *about target systems*. I turn to this question now.

One learns with a toy model if the toy model itself does not provide the learning but nevertheless plays a role in the learning process. I suggest viewing target-directed learning with a toy model as a three-stage process. In the first stage, a toy model is constructed by an epistemic agent to *capture* a pattern in a real world target $T$. This pattern can be a particular property of a target system or a relation between properties. In the second stage, toy models generate hypotheses about (types of) target systems. For a particular toy model $M$ it is the hypothesis $H$ that abstract properties (or relations

between abstract properties), which are instantiated by $M$, are instantiated in multiple target systems. In the third stage, an epistemic agent learns with the help of $M$ something about a target system $T^*$, which is *distinct* from the initial target system $T$, if the hypothesis $H$ is true for target system $T^*$, that is, the abstract property (or relations between abstract properties) is instantiated in target system $T^*$. The learning in this third stage consists in two insights: the agent learns that a target system of interest instantiates the abstract property and she learns, if she considers multiple target systems, the scope of the hypothesis $H$.[9]

Taking a step back from this account of learning reveals that the toy model plays the role of a hypothesis generator. The toy model itself cannot provide learning about the target system. The learning occurs, metaphorically speaking, in the moment in which the target system hits back in the expected way. Let me illustrate these rather abstract claims with the help of a few of my case studies.

The DY model captures, first, a key fact about income distributions, that is, that they are exponentially distributed. It comes equipped with the hypothesis that the abstract property of an exponential income distribution is instantiated in multiple target systems. These target systems are economies with different currencies, banking structures, and financial regulations. One learns something from the DY model about a so far unexamined real world economy, such as the economy of Denmark, if the income distribution of Denmark is exponentially distributed. If this condition is satisfied, one learns that the income distribution of Denmark is exponentially distributed. One learns something further, namely something about the scope of the initial hypothesis, by examining how many other economies also display an exponential income distribution.

The Kac ring captures, first, a key fact about the macroscopic behaviour of gases, that is, that they show an equilibrium like behaviour. The Kac

[9]I have given here sufficient conditions and not necessary conditions for learning with a toy model. The reason is that learning can also originate from the fact that patterns that one had reason to expect are *not* present in target systems. To keep the discussion manageable, I set this type of learning – one might want to call it "learning via discrepancy" – aside.

ring comes equipped with the hypothesis that there is a relations between the concepts of ergodic decomposition and equilibrium like behaviour, in the sense that systems that instantiate the former instantiate the latter. These target systems are fluids and gases with different micro constituents. One learns something from the Kac ring about a so far unexamined real gas if it is true that the gas with an ergodic decomposition shows an approach to equilibrium. If this condition is satisfied for a real gas, one learns that this gas shows an equilibrium like behaviour. One learns something further, namely something about the scope of the initial hypothesis, by examining how many other systems also instantiate this relation between the two abstract properties.

Schelling's checkerboard model captures, first, a key fact about distribution patterns within social groups, that is, that they are often segregated along binary lines (e.g., ethnicity in cities). The model comes equipped with the hypothesis that there is a relation between the concepts of mild preferences and a segregation pattern, in the sense that systems that instantiate the former instantiate the latter. These target systems are cities, schools, or universities with different social and geographic properties. One learns something from Schelling's checkerboard model about a so far unexamined particular real city if it is true that the city where people have mild discriminatory preferences shows segregation patterns. If this condition is satisfied for a real city, one learns that this city has segregation patterns. One learns something further, namely something about the scope of the initial hypothesis, by examining how many other systems also instantiate the relation between the two concepts.

Let me clarify my account of learning with toy models about target systems by highlighting some of its features. First, the discussed toy models offer different learnings on the third stage. The DY model comes equipped with the hypothesis that a single abstract property is instantiated in target systems. In contrast, the other two models come equipped with the hypothesis that a relation between two abstract properties is instantiated in target systems (e.g., the relation between ergodic decomposability and equilibrium like behaviour).

Second, one might ask whether my account is merely phenomenological. This would be the case if the learning provided by these toy models did

not classify as a type of explanation of the phenomena of interest. My account is phenomenological in nature. Even in the cases where the toy model comes equipped with a hypothesis stating that there is a relation between abstract properties, this is not a causal claim. Rather, the relation of abstract properties should be viewed as a claim about the co-instantiation of abstract properties in target systems. This relation of co-instantiation could also be described as a correlation. The natural follow-up question is, why it should be assumed that the postulated correlation between abstract properties should be expected in target systems different from the target system (or type of target system) that was used to recover the pattern in the first place.

In a sense, the move from a toy model result to the claim that the toy model result is instantiated in a variety of (type of) target systems bears similarity to Sugden's inductive leap. To recall, Sugden suggests that one can see model results as causal statements about target systems if there is a sufficient degree of similarity between the model and the target system. Note two crucial differences between my claim and Sugden's. First, I do *not* claim that toy model results should be read as causal statements. I only claim that, in some cases, toy model results suggest the presence of a correlation between properties in target systems (e.g., the correlation between mild preferences and a segregation pattern). Hence, my inductive leap is less ambitious than Sugden's regarding the nature of the postulated relation in a target system. Second, I claim that learning only occurs if the postulated relation is present in a target system. Sugden suggests that similarity judgements, and not the actual observation of causal relations in a target system, justify the inductive leap. Hence, my inductive leap is less ambitious in the sense that I do not require model-related grounds (such as similarity to a target system) that motivate the truth (or high likelihood) of a causal claim. However, I take it to be the case that even my less ambitious inductive leap requires some motivation: why should one expect a correlation of properties, postulated by a toy model, to hold in a variety of target systems?

In my view, the following analogy could provide some motivation for this expectation. Humans have developed a capacity for seeing by developing eyes. Eyes have been observed in organisms that differ significantly from

the human make-up. This fact motivates the assumption that eyes or eye-like systems can be expected in further organisms that differ significantly from humans. In parallel, there are patterns in the real world that can be recovered by toy models. These toy models are capable of adequately depicting these patterns *despite* the fact that the toy models look nothing like the system whose behaviour they capture. This fact motivates the assumption that other target systems, which could vary significantly from the initial target system, exhibit the same pattern. Let me emphasise again that, in my view, this inductive leap requires significantly less motivation than Sugden's inductive leap from a model result to the claim about a causal relation in a target system.

This allows me to state the relation between my account of learning based on toy models and knowledge about the causal structure of a target system in a precise manner: According to my account of learning, providing *causal* knowledge about a target system is not a necessary condition for learning with a toy model. However, gaining insights into the causal structure of a system – that is, if it turns out that the correlation, which is postulated by the toy model, is a causal relation in a target system – is compatible with my account of learning.

Third, my account does not presuppose that toy models come equipped with a well-specified domain of application. If this were the case, the second insight on stage three (i.e., to how many systems does the hypothesis apply to) would be defined ex ante, which is not always the case. The Kac ring illustrates this point: Although the key model result of the Kac ring consists in a relation between abstract properties, there remains an open question about what systems satisfy the antecedent condition (i.e., the ergodic decomposability). It must be shown that all measure-preserving systems satisfy this condition.

Fourth, one might object that my account does not respect the asymmetry of learning with a toy model: one learns something about a target system with the help of a toy model and not something about a toy model from an instantiating target system. In reply, I have accounted for this directionality of learning in the formulation of my account by introducing the epistemic

agent as the toy model user. In my view, an epistemic agent uses a model to realise some particular epistemic aims. It is this use that gives the learning a direction.[10]

Fifth, one might be worried that this account of learning with toy models about target systems does not adequately account for their status within the sciences. Put bluntly, are these models anything more than random hypothesis generators? If so, what distinguishes toy models from other sources of inspiration, such as gazing into an open fire? Let me say two things in response. To start, toy models are certainly distinct from these alternative forms of generating hypotheses since toy models allow an alternative form of learning. This was the conceptual learning highlighted in the previous section. Furthermore, it is true that my account moves toy models into the realm of discovery. Toy models themselves cannot justify their results. I see this as a by-product of my account that should be welcomed since it is in line with an empiricist account of knowledge generation.

This closes the exposition of my account of learning from and with toy models. I now turn to two arguments in its defence (Sections 5.4.2-5.4.3).

### 5.4.3 Learning and the hybrid representation condition

In the previous chapter, I argued that toy models satisfy the following hybrid representation condition:

**Hybrid representation condition** The basic entities, properties, and relations between the basic entities in the model do *not* represent features in any target system. However, there is a representational relation between model results, formulated on a more abstract level of description, and target systems.

The fact that toy models satisfy this condition does not undermine my claim that toy models can provide learning of the sort just described. To the contrary, my account of learning based on toy models can nicely be squared

[10]See the discussion of Lange (2014)'s objection against Batterman and Rice in Section 5.2.3.

with my characterisation of toy models. Let me elaborate on this with respect to the two distinct forms of learning.

With respect to learning from toy models, toy models can reveal inferential relations between concepts despite the fact that the basic entities and properties of these models do not represent features in any target system. The reason is simply that the inferential relations hold between the concepts itself. With respect to target-directed learning with toy models, toy models can play a part in revealing that an abstract property (or a relation between abstract properties) is instantiated in a target system and the scope of this claim for the same reason. For the learned facts do not make any reference to the non-referring entities or properties of toy models. Although the toy model is one system that instantiates the abstract properties (or their relations), and for this the basic entities and properties of the model matter, the learning with the model is not affected by the level of basic entities and properties since the learning is located on the level of abstract properties.

Take the DY model to start. One can learn that some developed economies show an exponential income distribution despite the fact that the money distribution principle in a collision of agents does not represent features of developed economies. From the Kac ring, one can learn that gases and fluids that have an ergodic decomposition show an equilibrium like behaviour despite the fact that the ring structure and the markers do not represent features in these gases and fluids. From Schelling's checkerboard model, one can learn that schools, universities, and cities where agents have mild discriminatory preferences show segregation despite the fact that the checkerboard structure and the movement rules of the model do not represent features in real world schools, universities, and cities.

### 5.4.4 Learning and lack of derivational robustness

An account of learning based on toy models is incomplete if it cannot account for a salient feature of this model class: the fact that many toy model results are derivationally non-robust, that is, the case that model results change if the assumptions of the toy models are altered. I engage with this observation

in two steps. First, I substantiate the point that many results of theoretical models are derivationally non-robust. Since I am discussing how toy models can be used in economic policy-making, I focus on economic models in this section. Second, I explore the impacts of a lack of robustness on the types of learning identified previously. To make this discussion as tangible as possible, I focus on Hotelling's model.

### Derivational robustness of economic models

Reiss (2008, p. 118) points out with respect to economics (italics are mine):

> (...) how sensitive to the precise setting many results in economic experiments are. An analogous fact is true, *and known to be true*, about results derived from *certain kinds* of economic models: they are, by and large, extremely sensitive to the precise set of assumptions made. Just change the transportation cost function slightly, change the setting from static to dynamic, introduce risk aversion, take out money illusion and so forth: all these factors will often make a dramatic difference to the derived result.

Let me focus on the two emphasised parts. To start, Reiss refers here to what he calls "mathematical models" (Reiss 2008, p. 113). In particular, he discusses Hotelling's model, which falls under my account of toy models. Furthermore, according to Reiss, the fact of the lack of derivational robustness of Hotelling's model can be generalised to other mathematical models (see Reiss 2013, p. 287; Hindriks 2013, p. 524). Indeed, Cartwright (2009, p. 50) makes a very similar point when she talks about the "paucity of economic principles". Because there are not many well-established economic principles that can be included in an economic model, one needs a rich set of over-constraining structural assumptions. As it turns out, the model derivations depend heavily on these structural assumptions (Cartwright 2009, p. 53).[11] This can be viewed as one way of spelling out her earlier comments about

[11] To avoid confusion at this stage, Cartwright uses the term "structural assumptions" differently to Kuorikoski et al. (2010) (see Section 1.3.1). Cartwright's structural assumptions are Kuorikoski et al. (2010)'s tractability assumptions.

the specific assumptions that are needed to obtain the results in game theory (Cartwright 1999, p. 148-149).

This problem with respect to economics was probably stated in its most general form by McCloskey (1993, p. 235) as the A-Prime/C-Prime Theorem:

> For each and every set of assumptions A implying a conclusion C, there exists a set of alternative assumptions, A', arbitrarily close to A, such that A' implies an alternative conclusion, C', arbitrarily far from C.

What does this lack of robustness mean for the type of learning defended in this chapter?[12] To answer this question, I turn to a close examination of the derivational robustness of Hotelling's model.

### Analysing a case: Hotelling's model

In this section, I draw from the description of Hotelling's model that I provided in Section 4.2.4 to argue for the claim that the lack of robustness does not rule out learning based on toy models. Is the key insight from Hotelling's model, that is, the principle of minimal differentiation, derivationally robust? As the discussion will make clear, this is not an all-or-nothing matter but a matter of degree: The principle of minimal differentiation exhibits different stability properties given different types of changes in the model assumptions.

Hotelling devotes the last two pages of his article to the question of derivational robustness. He considers the following changes to his model: 1) buyers not uniformly distributed along the line but distributed with varying density, 2) buyers not distributed along a line but on a two-dimensional plane, 3) more than two sellers, 4) different transportation cost functions (i.e., not linear in distance), 5) additional factors that are relevant to a decision of a buyer (e.g., product features, such as quality), 6) introducing a seller's ability to discriminate between prices, and 7) changing the elasticity of demand

[12] This lack of robustness of mathematical model results is not only observable in economics. As Weisberg and Reisman (2008, p. 118)'s discussion makes clear, the Lotka-Volterra model exhibits a variety of non-robust model results. For example, introducing a carrying capacity of the ecosystem for prey in a structural derivational robustness analysis destroys the property of undampened oscillation of the predator/prey population.

(Hotelling 1929, p. 56). For all these changes, he states that the key model result, that is, the principle of minimal differentiation, holds (Hotelling 1929, p. 56). To be precise, regarding changing the elasticity of demand, Hotelling (1929, p. 56) postulates a slight difference in the result compared to his initial analysis, that is, the tendency of a seller to establish its business "excessively close to (...) [a competitor] is less marked".

Interestingly, Hotelling does not offer any calculations for these robustness claims with respect to these seven changes (see also Reiss 2012, p. 52). Even more interestingly, the robustness claims do not withstand scrutiny, as is carefully summarised by Brenner (2001). Brenner (2001, p. 20) ends his survey article on the research tradition started by Hotelling's model as follows: "This survey reveals that differentiation – either in geographic or in product space – depends delicately on parameters of the market structure".

Recall Hotelling's principle of minimal differentiation: It says that sellers cluster together (or products show very similar characteristics) in markets with competing sellers. I focus on three changes under which this result is derivationally non-robust.[13]

First, if one solely changes the assumption about the number of suppliers in the model (i.e., from two to three), the suppliers still have an incentive to concentrate in the market centre, however, there is no location equilibrium (Brenner 2001, p. 17). Second, if one changes the assumption that consumers are price inelastic (i.e., they do not change their demand given a decrease or increase in the price of the goods), and makes some assumptions about the consumers' reservation price, one can observe any location pattern between minimal and maximal differentiation (Brenner 2001, p. 14-15). Third, if one changes the transportation cost function (Hotelling assumed a function that is linear in the distance travelled with the purchased good) to a function $d(x) = bx^a$ (where $d(x)$ denotes the costs and $b > 0$, $1 \leq a \leq 2$), then one observes in-between to maximum differentiation.

Now, what are the implications of these observations for the ability to

[13]Brenner (2001) discusses additional changes to the basic set-up of Hotelling's model, including changing customer distributions, introducing uncertainty about product characteristics, the opportunity for firms to collude on prices, and different pricing strategies.

learn something from Hotelling's model? Let me break this question down into two parts: First, what is the impact of the lack of derivational robustness of Hotelling's main claim for *learning from* Hotelling's model? Second, what is the impact of this observation on *learning with* Hotelling's model about target systems?

One can distinguish between, what I will call, *unproblematic* and *problematic* forms of a lack of derivational robustness.[14] A lack of derivational robustness is unproblematic if the new model result $R'$ instantiates the same abstract property (or relation between abstract properties) that was instantiated by the initial model result $R$. A lack of derivational robustness is problematic if the new model result $R'$ does not instantiate the same abstract property (or relation between abstract properties) like $R$. It should be straightforward why the first type of lack of robustness is unproblematic. If the same abstract property (or relation between abstract properties) is instantiated, then no change has taken place on the relevant level of learning. The model still comes equipped with the same hypothesis and, hence, learning can take place if the hypothesis is true of a particular target system (or with respect to the scope of the hypothesis). This can be shown with the help of Hotelling's model. Changing the number of suppliers in the model (from two to three) reveals that there is still a tendency to agglomerate in the centre. Importantly, this model result $R'$ also instantiates the abstract property of similarity in location or product characteristics. Hence, the relations between abstract properties that were postulated by Hotelling's initial model remain unchanged.

However, there are also examples of a problematic lack of derivational robustness with respect to Hotelling's model. If one changes the assumption that consumers are price inelastic, one can observe any location pattern between minimal and maximal differentiation (Brenner 2001, p. 14-15). Or, if one changes the transportation cost function to a power function, then one observes in-between to maximum differentiation. In both cases, the relation between the abstract properties of competition and minimal differentiation

[14]Note that I use the notion of 'problematic' here in the sense of prima facie problematic for my account of learning from and with toy models.

no longer holds (or is no longer the sole outcome, as it was in the first of the two cases). This allows to pin down the pressing question more precisely: In what sense, if at all, can there be learning from and with a toy model result that is *derivationally non-robust in a problematic sense*?

With respect to learning from Hotelling's model, it is worth noting that inferential relations between concepts are revealed even in the case of problematic derivationally non-robust results. To start, Hotelling's initial result reveals a relation between profit-maximising behaviour on the supplier side, preference satisfaction on the demand side, duopolistic market structures, and sets of possible actions (e.g., spatial movement) on the supplier and demand side. Furthermore, the lack of derivational robustness in a problematic sense shows that these inferential relations can change substantially. But the outcomes of these changes reveal new inferential connections. For example, the changes in the price elasticity of the demand side or the transportation costs reveal a new inferential relation as specified above. This point can be metaphorically put as follows: instead of viewing Hotelling's toy model as a fixed set of assumptions that reveals one insight (i.e., Hotelling's principle of minimal differentiation), the model should be viewed as a web of concepts that can reveal different insights (i.e., inferential relations) given different configurations of assumptions. Hence, what one gains from Hotelling, even in these problematic cases of lack of derivational robustness, is a particular kind of conceptual learning.

One can object to this view by pointing out that the alternative inferential relations, which are revealed in a problematic case of a lack of derivational robustness, are not interesting, since they mainly point out the role of tractability assumptions. Recall that tractability assumptions are those that are empirically not well (or not at all) motivated but are facilitating the derivation of results (Kuorikoski et al. 2010, p. 547). I do not think that this objection is convincing. First, as my discussion of Hotelling's model makes clear, the changes made in the robustness analysis reflect substantial conceptual changes. Second, and more importantly, assumptions can be classified differently given different epistemic contexts.

With respect to learning with Hotelling's model, I think the learning is

rather limited. The fact that its key model result is non-robust leaves two types of insights about target systems in place. First, Hotelling's model still provides a qualitative accurate description of a variety of target systems, that is, the presence of a pattern of spatial concentration or degree of similarity in an offering given a competitive structure. It is true that – given the lack of robustness – Hotelling's model provides this qualitative accurate description only under a specific set of assumptions. Note that the smaller the set of assumptions under which the model yields the qualitatively accurate description of the target system, the more fragile, so to speak, is the model's descriptive accuracy, and, hence, the less value one should attribute to this descriptive function of a toy model.

Second, Hotelling's model still shows that a particular modelling technique can be applied to a particular type of social phenomenon, that is, game theoretic modelling to strategic social interactions. This insight from a toy model is even more indirect than my account of learning with toy models already suggests. The toy model introduces or exhibits the applicability of a particular modelling technique to a particular problem or type of target system. Thereby, the toy model provides an insight about the tools that can be brought into play with respect to a particular problem or target system. Exploring these modelling techniques can yield more refined models that can provide predictions or explanations that are more accurate. Let me briefly leave my analysis of Hotelling's model to further illustrate this point. Consider Schelling's checkerboard model. This model is widely regarded as the starting point of applying evolutionary game theoretic modelling to social phenomena (see, for example, Aydinonat 2006). By now, there exists a large body of literature that uses evolutionary game theory to model social phenomena (see, for example, Alexander 2007 who applies evolutionary game theory to questions about the foundations of morality). However, note that Schelling's key model result (i.e., the segregation pattern) has different robustness properties than Hotelling's principle of minimal differentiation. The segregation pattern emerges also if one changes the size of the neighbourhood and the individual preferences (Muldoon et al. 2012). In my view, the fact that a model result is derivationally robust is not a necessary condition for the aforementioned

introduction of a modelling technique. To see this, note that one can distinguish between two questions: 1) Can one prima facie apply a modelling technique to a particular question or (type of) target system? 2) If one can apply it, is the application fruitful, for example in the sense that it leads to robust model results? In the case of Hotelling's model, in contrast to the case of Schelling's model, the answers to the second question differ. However, this does not rule out the insight one can gain from Hotelling's model regarding the first question.

This concludes my analysis of Hotelling's model. I have shown that there needs to be a fine-grained assessment of the lack of derivational robustness of Hotelling's principle of minimal differentiation. The discussion revealed that two distinct types of learning are still possible in this case: a form of conceptual learning from the toy model and the introduction of a potentially valuable modelling technique.

## 5.5 The use of toy models in economic policy-making

In the previous section, I introduced my account of learning from and with toy models and supported it by articulating two of its features. With this in place, I can return to the key question of the chapter: What epistemic role can toy models play in economic policy-making? The discussion so far supports two roles for toy models: a conceptual clarification role and a descriptive role in relation to target systems.

A conceptual clarification role follows directly from my analysis of learning from a toy model. I argued that toy models by themselves can reveal inferential relations between concepts. This is the case both for derivationally robust and derivationally non-robust results. This kind of conceptual learning can be relevant in those policy-making contexts in which conceptual clarifications are needed. Such a situation is, for example, present if two or more actors involved in a policy-discussion disagree about the importance of certain assumptions since they disagree about their implications. In such a

context, a toy model can reveal in a formally precise way the implications of these debated assumptions. However, this is not the key issue that economic policy-makers face. Rather, disputes about concrete policy issues (e.g., What should be the price for carbon dioxide certificates? Should Heathrow or Gatwick Airport receive the right to build an additional runway?) revolve around questions as to what the impacts of particular policy interventions may be. This shifts our focus of attention into the realm of prediction, explanation, and control. Since these are target-directed epistemic aims, let me turn to what can be gained from toy models in relation to target systems.

Here, a descriptive role of toy models in relation to target systems comes into play. To start, consider the cases of robust toy model results or unproblematic derivationally non-robust results. Here, the toy model serves as a tool that suggests instantiations or co-instantiations of abstract properties in a target system. For example, in the offline retail market for electronic consumer goods in London there is a local clustering. However, as pointed out above, the actual empirical verification of this claim (i.e., observing the local clustering of sellers of electronic goods) is the learning step. Hence, toy models should not be seen as supplying predictions that turn out to be correct with a high probability or to reveal stable causal relations that can be exploited for policy interventions. The reason for this can now be captured succinctly: Learning with toy models means that the toy model is a tool for hypothesis generation. This tool supplies hypotheses of a particular kind: claims that abstract properties (or relations between abstract properties) are instantiated in multiple target systems. However, in the case of economic policy-making one is interested in a particular target system. Here, information about abstract properties across diverse target systems is of less importance because the details of the particular target system matter.

Next, consider the case of problematically derivationally non-robust toy model results. The discussion in the previous section has made clear that even in this case, there are some insights to be gained from a toy model in relation to (types of) target systems: the toy model result still captures qualitatively a feature or relation in a target system (albeit only under a restricted set of assumptions) and can introduce a fruitful modelling technique that might

lead to the development of models that allow for quantitatively accurate predictions or yield explanatory insights. As before, these two insights that can be gained directly with the toy model are likely to be only of minor relevance for economic policy-making. With regards to the descriptive accuracy about qualitative features of a system, toy models do not provide detailed knowledge about the causal structure of a target system that could be used for prediction and control. With regards to the the introduction of modelling techniques, the toy model itself serves only as the starting point of a model construction process, that, ultimately, might lead to a more refined model of a target system that can be used for economic policy-making.

Taking a step back reveals that toy models can only play a limited role in economic policy-making. There is a clear conceptual role for toy models, albeit the conceptual clarification tasks in economic policy-making are, arguably, limited. There is a descriptive role for toy models, albeit this descriptive role does not supply the information needed for explanation and control regarding particular target systems of interest. This upshot should caution against reliance on these models in economic policy-making. Or, to put it positively, economic policy-makers should be aware for what epistemic purposes a toy model is used.

## 5.6 Concluding remarks

In this chapter, I have looked at the epistemic value of toy models to address my third research question: What should one do if model results are non-robust, that is, model derivations are not in agreement? I argued that non-robust toy models still can provide conceptual learning or can introduce a potentially valuable modelling technique. However, I have emphasized that these two types of insights are only of limited use for economic policy-making: toy models do not reveal stable, causal relations that can be exploited for interventions in particular target systems.

My account goes some way towards providing a justification for why toy models with non-robust results are so prevalent in scientific disciplines. Importantly, my account suggests an epistemic reason for why this might be

the case. Using toy models allows, potentially successfully, the depiction of macro-scale behaviour across diverse target systems, can offer some conceptual insights by uncovering inferential relations, or introduce potentially fruitful modelling techniques.

This chapter concludes my engagement with the application of robustness considerations to policy-making domains. In the final chapter, I ask what one should do if there is a lack of measurement robustness and, given the situation one is in, if one is forced to choose between models or theories.

# Chapter 6

# What if Robustness Analysis Fails? An Account of Theory and Model Choice

## 6.1 Introduction

As I pointed out in the introductory survey of the literature on robustness analysis, there is an open question about what to do if the robustness condition is not satisfied. This can happen with respect to derivational robustness, that is, the set of considered models does not yield the same (or reasonably similar) outcomes; or it can happen with respect to measurement robustness, that is, different types of evidence support conflicting hypotheses. Here, I focus on the latter situation and address my fifth and final research question: What should one do if one faces a situation of non-robust evidence from multiple evidential sources?

One natural response to this epistemic predicament is to suspend one's judgement about hypotheses for which evidence is not robust. Suspending judgement is compatible with pursuing additional investigations, which can lead to robust evidence. However, one might not always be in the situation that one can suspend one's judgement. Consider the following hypothetical situation.

A policy-maker faces the task of safeguarding a city area against flooding hazards from a nearby river. She must decide between investing into a new technology for flood protection (a sophisticated drainage-pump system) or setting aside money to build flood protection walls. The flooding hazard is expected to materialise in 15 to 20 years from now. The decision must be taken now, since the development of the new technology requires time. Only one of the two actions can be chosen since there are not enough resources to pursue both measures. Two possible states of the world are critical for the decision the policy maker faces: whether the river level will be above or below 4 metres in 15 to 20 years from now. She draws a decision matrix (see Table 6.1).

| | River level $\geq$ 4 metres | River level $<$ 4 metres |
|---|---|---|
| Invest in new technology | 100 | -100 |
| Set money aside for walls | -100 | 100 |

Table 6.1: The decision matrix of the policy-maker

If the river level will be equal to or greater than 4 metres, investing into the new technology is optimal since only the new system will protect the city. If the level will be below 4 metres, setting money aside for building walls is optimal.[1] Now assume further that the evidence about future river levels is inconclusive, in the sense that different types of measurements and model predications are not in agreement. In fact, due to the specific geological situation of the region, the scientists providing the evidence are unable to give precise[2] or imprecise probabilities[3] over these two future states of the world.

[1]Let me expand on the interpretation of the utility numbers in the table. One way of reading them is in terms of a relative comparison. That is, "walls" are better than "new technology" when water levels are low, but "new technology" is better than "walls" when water levels are high. Note that the precise numbers do not matter for my argument. It only matters that one has a symmetric decision matrix, as I outline below.

[2]Ascribing precise probabilities to the two states in this examples requires a single probability density function. Such an ascription could be for example: P(River level $\geq$ 4 metres) = 25% and P(River level $<$ 4 metres) = 75%.

[3]Ascribing imprecise probabilities to the two states in this examples does not require a single probability density function. Instead, one could rely on a set of such functions. If one goes imprecise, an ascription could be the following: P(River level $\geq$ 4 metres) = (25-45%) and P(River level $<$ 4 metres) = (55-75%).

Since no precise or imprecise probabilistic information is available, standard decision theoretic tools fail to help here. Without probabilistic information one cannot perform standard expected value analysis.[4] Since the decision matrix is symmetric, going for the decision criterion of maximising the value of the worst potential outcome (i.e., maximin) also does not yield a resolution here. This leaves one with two options: being indifferent between the two options and, in the light of the need of choosing a measure, one flips a coin; or to make an informed choice between the measures. Since one is in a policy-making context in which accountability of the policy-maker to the general public is likely, I think opting for an informed choice is the more realistic scenario. To make this informed choice between the measures, it is reasonable to go back to the evidence set and make a choice between the competing models that support the different predictions about the future river levels.

Let me be clear about the status of this example. It should not show that going back to the evidence set and making a choice is the only compelling or rational permissible action for the policy maker. Nor do I want to imply that I have given an exhaustive overview of decision rules in situations where precise or imprecise probabilistic information is missing. The upshot of this example is rather that there are some scenarios in which choosing between models or theories that are part of or underlie an inconclusive evidence set is a reasonable step to take. The example allows identifying some characteristics of these scenarios: it is a decision situation in which a) alternative options have radically different outcomes given different states of the world, b) one cannot suspend judgement due to the time sensitivity of the choice, c) there is sparse information about the decision-relevant states of the world, which prohibits (or at least makes it difficult) to use standard decision theoretic tools, and d) the evidence, which could help resolve the factual issue about what the relevant future state of the world is for the decision maker, consists

[4]Standard expected value of an option (e.g., investing in new technology) is calculated by multiplying the utility of an outcome with its probability of occurrence for the different possible states of the world. If one has imprecise probabilistic information, one could calculate expected values of options for different probability functions within the set that is given by the evidential constraints (for more on this see Chapter 3 of Resnik 2011 and Troffaes 2007).

of or is underlied by models or theories.

Having said this, let me argue in more detail why a choice between models or theories should be a reasonable step to take. I motivate this step by highlighting the shortcomings of the two main alternative ways of proceeding. Instead of choosing between the models or theories, one could endorse them simultaneously. However, this would not solve the problem, since one would be back in the situation, per assumption of the example, in which different types of evidence support conflicting hypotheses. Alternatively, one could also try to ascribe weights to the different models or theories, and, then, use these weights to arrive at a weighted value for the predictions entailed by the different models or theories. This weighted value of a prediction could then be the basis for choosing between the different actions. Consider Figure 6.1 for a simplified visualisation of the situation.

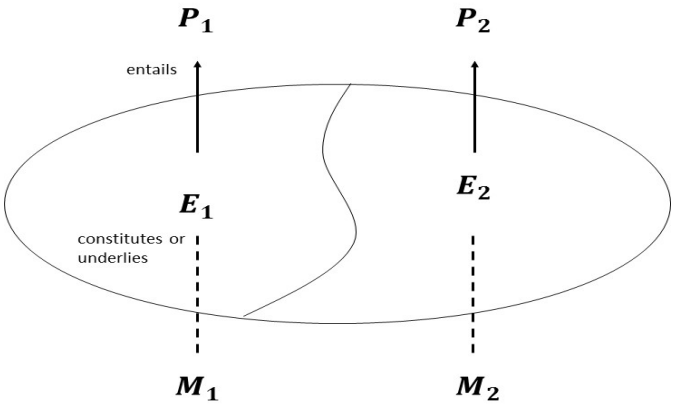


Figure 6.1: Evidence set with two elements $E_1$ and $E_2$ that support two conflicting predictions $P_1$ and $P_2$. The evidence set is constituted or underlied by two competing models or theories $M_1$ and $M_2$.

I think this approach is likely to run into two problems. First, it might be that the predictions $P_1$ and $P_2$, which are entailed by the models or the corresponding pieces of evidence, are such that an averaging into an overall number is not possible. Consider our example again. If the two parts of the evidence set support the predictions "river level above or equal to 4 metres"

and "river level below 4 metres", then these two statements cannot be averaged. Averaging presupposes predictions stated in precise numerical values. Second, and more importantly in my view, it is not clear that one has enough information to ascribe precise weights to the models or theories under consideration. Recall that I assume in the example that one must be able to provide some sort of justification for the policy option one chooses in the end, and, hence, resorting to subjective degrees of belief for ascribing these weights is problematic.

How does one rationally choose, then, between different models or theories? A look at the history of science reveals many such choice situations: for example, the choices between the geocentric and heliocentric model of the solar system, between phlogiston or oxygen based accounts of combustion, and between Newtonian and Einsteinian dynamics. Kuhn (1977)'s paper still provides the relevant stage-setting for discussing theory choice. Kuhn claimed that a variety of epistemic values, most importantly scope, fruitfulness, accuracy, simplicity, and consistency, influence theory choice. However, he rejected the idea that there is a unique algorithm to aggregate the information provided by these values (Kuhn 1977, p. 322, 326).

Recently, Okasha (2011) gave a new twist to the debate of theory or model choice.[5] He proposed an analogy between social choice and theory choice. If the analogy holds, an Arrovian impossibility result emerges: there exists no aggregation procedure that yields a complete and transitive ranking of the alternative theories considered and that satisfies a set of intuitively compelling conditions. This result is troubling since it casts doubt on widely used multi-criteria aggregation procedures for assessing the suitability of a theory.

In this chapter, I provide a procedure for theory choice that allows one to perform this choice in a rationally defensible way. I argue that Okasha's analogy between social choice and theory choice does not hold since the theory choice problem can be viewed in a cardinal context. Theory choice in sciences is better described as a weighing processes than in the framework of ordinal social choice theory, which emphasises consistency. My starting point

[5]To simplify the exposition, I use subsequently solely the expression 'theory choice'. Choices between models are always subsumed under this notion.

is a suggestion made by Okasha in the latter part of his paper, namely to enrich the informational basis of the analysis to allow for inter-criteria comparability. Okasha refers to Sen (1977, 1986) who argued that some degree of interpersonal comparability is needed in order to avoid Arrow's impossibility result. This shift in the problem description allows me to use a tool that has been successfully applied in the social choice context, that is, scoring rules, to make the notion of weighting in the context of theory choice more precise. I argue that a general scoring rule characterised by Gaertner and Xu (2012) is flexible enough to illuminate, and solve, the problem of theory choice.

The chapter is structured as follows. First, I briefly discuss Kuhn's treatment of theory choice (*Section 6.2*). Second, I introduce Okasha's reconstruction of Kuhn and his impossibility result (*Section 6.3*). Third, I discuss the reactions in the literature to Okasha's claim and locate my reply strategy against this background (*Section 6.4*). Fourth, I introduce my claim that Okasha's analogy does not hold and substantiate this by providing a procedure by which the aggregation process in theory choice can be performed (*Section 6.5*). Fourth, I explain what my solution reveals about the aggregation across different scientists (*Section 6.6*).

## 6.2 Kuhn's discussion of theory choice

In *The Structure of Scientific Revolutions*, Kuhn lays out a picture of the development of mature scientific disciplines that is characterised by two phases: a normal scientific phase in which the scientific community acts based on a shared paradigm that defines exemplars of problem solutions and a revolutionary phase in which multiple paradigms compete (Kuhn 1962). The phase in which both the Ptolemaic and the Copernican theory of the solar system where used by different groups of astronomers is such a revolutionary phase. This picture of the development pattern of scientific fields prompts a question: How are choices made between these competing paradigms?

Significant parts of Kuhn (1962) are devoted to the question whether competing paradigms can be compared at all. One of the most controversial claims of Kuhn is that competing paradigms express incommensurable world views,

that is, there is no common scale on which two paradigms can be compared (Kuhn 1962, p. 147-150). Moving between incommensurable paradigms is described by Kuhn as a "switch in visual gestalt", which prompts scientists to see the world differently (Kuhn 1962, p. 111). Kuhn also argued that this gestalt switch does not happen solely due to epistemic reasons that speak in favour of a new paradigm but also due to non-epistemic reasons, such as the preference for the beauty of a theory (Kuhn 1962, p. 144, 151).

In a later article entitled *Objectivity, Value Judgement, and Theory Choice* Kuhn returns to the question of theory choice. He attempts to pin down more precisely in what way epistemic and non-epistemic factors influence theory choice. Kuhn argues that at least five values characterise a good scientific theory. First, a good scientific theory should be *accurate*, that is, its consequences should be in agreement with observational results. Second, a good scientific theory should be *consistent*, that is, it should not contain logical contradictions and be compatible with other scientific theories. Third, a good scientific theory should have *broad scope*, that is, it should have a domain of applicability that exceeds the phenomena the theory was initially supposed to explain or predict. Fourth, a good scientific theory should be *simple* in the sense that it unifies a broad set of phenomena under a few theoretical assumptions. Fifth, a good scientific theory should be *fruitful*, that is, it should disclose new phenomena or new relationships among already known phenomena (Kuhn 1977, p. 321-322).

With respect to these values, I take Kuhn's key claim for my discussion to be the following: Given two (or more theories) and the epistemic values of accuracy, consistency, scope, simplicity, and fruitfulness, different scientists can arrive at more than one ranking of the alternatives even if they agree that the evaluation should be done solely with reference to these epistemic values. This corresponds to saying that there is more than one algorithm to determine the overall ranking of the alternative theories based on these epistemic values (Kuhn 1977, p. 322, 326).

This formulation entails two important clarifications regarding what I take to be the problem of theory choice. To start, I am setting aside two arguments in Kuhn's earlier writings. The first argument can be put as follows: Theories,

more precisely paradigms, come equipped with standards for assessing theories. These standards can vary across different paradigms. Hence, there is no unique way of choosing between paradigms (Kuhn 1962, p. 6, 141). The second argument goes like this: Even if there are shared standards for assessing paradigms, paradigms do not solve an identical set of problems. Accordingly, if one chooses between two or more paradigms, one must weigh the importance of the problems against each other. Hence, there is no unique way of choosing between paradigms (Kuhn 1962, p. 85, 103; Hoyningen-Huene 1993, p. 242). Consequently, I only deal with the situation in which the set of problems as well as the standards of theory evaluation are shared.[6] If these arguments are set aside, are there further reasons that Kuhn puts forward in defence of the non-algorithmic nature of theory choice?

Kuhn provides the following two additional arguments. To start, epistemic values can be interpreted differently by scientists who are involved in an evaluation process (Kuhn 1977, p. 322). Hence, it is possible that two scientists, who are committed to the same set of epistemic values, come up with different rankings of the alternatives under consideration (Hoyningen-Huene 1993, p. 236). Furthermore, the epistemic values can be weighted differently by scientists who are evaluating theories (Kuhn 1977, p. 322; Hoyningen-Huene 1993, p. 236). The commitment to the same set of values does not entail or presuppose a commitment about their relative weight. Hence, it is again possible that two scientists, who are committed to the same set of epistemic values, come up with different rankings of the alternatives under consideration.

Furthermore, my formulation entails a particular view regarding the level at which the problem of theory choice is located. To be more precise, is the theory choice problem an issue on the level of individual scientists or the scientific community? This is a tricky question since Kuhn's remarks on this give a mixed picture:

> (...) it is the community of specialists rather than its individual members that makes the effective decision. (Kuhn 1962, p. 200)

[6] I take it to be the case that Okasha (2011), as well as the replies in the literature, which will be discussed below, share this focus. It is important to note that thereby Kuhn's discussion of methodological incommensurability is left aside.

> (...) shared values can be important determinants of *group behaviour* even though the *members of the group* do not all apply them in the same way. (Kuhn 1962, p. 186, my emphasis)

How should one understand these remarks about decision making on the group level and how are they related to the decision making of individual scientists?

I follow Hoyningen-Huene (1993)'s interpretation here. According to him, the relevant debate occurs on the level of the individual scientist. The choice of an individual scientist is influenced but not determined by epistemic values. The individual scientist chooses a preferred theory and invests her energy in its development. Which theory comes out on top in the scientific community is determined by a historical process consisting of the choices of individual scientists (Hoyningen-Huene 1993, p. 153-154). Subsequently, I first focus on the choice situation of an individual scientist (*Sections 6.4-6.5*). Based on my procedure for this situation, I then discuss the aggregation across different scientists (*Section 6.6*).[7]

## 6.3 Okasha's Arrovian reconstruction of Kuhn

Okasha treats each of Kuhn's epistemic values (simplicity, accuracy, fruitfulness, consistency, and scope) as if it were an individual with a preference ordering over the alternative theories. To be more precise, every epistemic value can be viewed as a decision criterion $n \in N$ (where $N$ is the set of relevant criteria) that can be expressed as a binary relation, $R_n$ (e.g., "is at least as simple as", "has at least the scope as"), defined on the set of alternative theories $X$. Each binary relation, $R_n$ imposes a weak ordering on $X$: an ordering that is reflexive, transitive, and complete (Okasha 2011, p. 91).

Given this framework, Kuhn's algorithm can be expressed as a theory choice rule (Okasha 2011, p. 92). A theory choice rule is a mapping from the

[7]Note that Weber (2011) offers an alternative interpretation of Kuhn. He views Kuhn as a social epistemologist (Weber 2011, p. 3), who treats the scientific community level as the relevant decision making entity (Weber 2011, p. 7). Accordingly, Weber would reject my two step approach to the choice problem.

set of all logically possible combinations of weak orderings $(R_1, ..., R_n, ..., R_t)$ to a single weak ordering $R^*$ which is the aggregate relation defined on the set of alternative theories $X$ and interpreted as 'is at least as good as'.

According to Okasha, all five requirements that Arrow (1963) postulated in his work on the non-existence of a social welfare function have to be met by a theory choice rule (Okasha 2011, p. 92). To start, the aggregate relationship has to be a weak ordering, that is, the aggregate relationship has to be transitive and complete. This is a requirement of collective rationality. The postulate of an unrestricted domain (**U**) means that the theory choice rule should yield an overall ranking $R^*$ for all logically possible combinations of $t$-tuples of binary relations $R_n$ (Okasha 2011, p. 92). The requirement of the weak Pareto principle (**P**) states that if theory $T_1$ does better than another theory $T_2$ with respect to all considered epistemic values $1, ..., n, ..., t$, then $T_1$ should be preferred to $T_2$ overall (Okasha 2011, p. 92). Arrow's non-dictatorship (**D**) requirement states that there is no epistemic value such that if this value ranks, for all profiles of preference rankings, any $T_1$ above any other $T_2$, $T_1$ is ranked automatically above $T_2$ in the overall ranking (Okasha 2011, p. 93). Finally, the independence of irrelevant alternatives (**I**) condition requires that the overall ranking of any $T_1$ and $T_2$ depends only on how the epistemic values rank $T_1$ and $T_2$ and not on how they rank other theories in relation to $T_1$ and $T_2$ (Okasha 2011, p. 93).

Given this reconstruction of the theory choice problem, Arrow's famous impossibility result applies: For a finite number of epistemic values and at least three alternative theories, there exists no theory choice function satisfying conditions **U**, **P**, **D**, and **I** (Okasha 2011, p. 93).

## 6.4 Responses to Okasha

Okasha's paper has stimulated a variety of responses. The responses fall into two broad categories: those who argue that Okasha's description of the theory choice situation is misguided and those who try to avoid the Arrovian impossibility result by showing that at least one of Arrow's conditions cannot be applied in the case of theory choice. Below, I briefly discuss the key

contributions and highlight some shortcomings of them. Importantly, these shortcomings should not be read as knock-down arguments. Instead, they should motivate the exploration of an alternative way of dealing with Okasha's impossibility result.[8]

### 6.4.1 The appropriateness of Okasha description of the problem

A first potential escape route is to point out that Arrow's impossibility result requires three or more alternatives whereas theory choice problems are usually binary. Kuhn's own examples such as the choice between geocentrism and heliocentrism or between the phlogiston and oxygen theory of combustion involve only two alternatives. This escape route is brought up and discussed by Okasha (Okasha 2011, p. 477-478).

I think Okasha is right in pointing out that this escape route is a dead end. For example, in cases of model selection in climate science, a large number of models is compared and evaluated. In cases of establishing an explanation for a correlation between two variables $x$ and $y$, multiple hypotheses need to evaluated (i.e., $x$ causes $y$, $y$ causes $x$, and $x$ and $y$ are the effect of a common cause) (Okasha 2011, p. 478). So, relevant choices in scientific practice likely involve three or more alternatives.

Marcoci and Nguyen (2017) argue that Okasha describes the rationality of science wrongly in an all or nothing manner. Scientific rationality has to be seen as a graded notion allowing for rationality in degrees (Marcoci and Nguyen 2017, p. 322). The basic idea is to ask whether an aggregation function that satisfies the Arrovian conditions of weak Pareto, independence of irrelevant alternatives, non dictatorship and universal domain is normatively acceptable on a profile by profile basis. An aggregation rule is normatively acceptable with respect to a set of profiles if and only if it yields a transitive and complete overall ranking (Marcoci and Nguyen 2017, p. 324). This way

[8]In this review, I do not include the papers that discuss the Arrovian impossibility result with respect to the confirmation of empirical hypothesis. See Stegenga (2013), Lehtinen (2013), and Cresto et al. (2017) for a discussion.

of thinking about the problem opens up the possibility to argue that some aggregation rules are normatively compelling for some profiles and yet not acceptable for others. With this concept of "minimal rationality" Marcoci and Nguyen want to express the idea that the number of profiles for which a particular aggregation functions yields a normatively acceptable outcome is key for assessing the rationality of this particular aggregation function (Marcoci and Nguyen 2017, p. 326). They show that the number of alternatives and theoretical virtues considered in a choice problem are crucial for the degree of rationality of an aggregation function. For example, given three alternatives and five epistemic values, pairwise majority rule yields an intransitive overall ranking for only 7% of the possible profile combinations (Marcoci and Nguyen 2017, p. 328).

Marcoci and Nguyen (2017)'s proposal is interesting since it opens up an alternative way of thinking about the rationality of science. In my view, one question needs, however, to be addressed. As Marcoci and Nguyen (2017, p. 329-330) point out, the question of the threshold value at which it is warranted to use a particular aggregation procedure remains open. Crucially, the merit of their proposal depends on being able – at least in principle – to give such a threshold value. It is not clear to me whether this should be, and, if so, can be done in a context-depend manner.

### 6.4.2 The applicability of the Arrovian conditions

If one accepts Okasha's problem description then one can solely get rid of the Arrovian impossibility result by denying the applicability of one (or more) of Arrow's conditions in the case of theory choice.

#### The non-dictatorship condition

Weber (2011) argues that the non-dictatorship condition (**D**) must be given up in the context of theory choice. According to Weber, fruitfulness must be viewed as a dictatorial criterion amongst the epistemic values (Weber 2011, p. 6). Weber starts by pointing out that there is no reason why science should be committed to weight all epistemic values equally. Although individual

scientists' choices are influenced by multiple (or all of) the five epistemic values, the choice on the community level of a paradigm is determined by the fruitfulness of the paradigms under consideration (Weber 2011, p. 7-8).

I do not think that Weber's escape route is compelling, since, as I have pointed out above, I see the theory choice problem as a two-step process. In a first step, scientists make their choices and, then, the aggregate of these choices leads to a result on the community level. I think it is implausible to assume that there is an additional, ultimately relevant, mechanism that solely operates on the community level.

### The completeness condition

Bradley (2016) argues that the completeness condition, which is part of the collective rationality condition regarding the all-things considered ranking of alternatives, must be given up. He suggests viewing the epistemic values as criteria that constrain theory choice but that do not determine its outcome. Importantly, he does not see this as a shortcoming of scientific rationality but as a natural feature of rationality more generally since rationality "'can be silent" (Bradley 2016, p. 6). To see this, consider a bet on a fair coin toss. Since the probability of heads and tails coming up is exactly equal, rationality does not determine what one should bet. However, rationality is still constraining the acceptable bets since it would be irrational to accept a bet on heads at shorter odds if a bet on heads at longer odds were available (Bradley 2016, p. 6). In the case of theory choice, the epistemic values provide rational constraints in the form of non-complete orders of the alternatives under consideration (Bradley 2016, p. 2). The key constraint that Bradley discusses is provided by the weak Pareto condition (**P**) that states that if theory $T_1$ does better than theory $T_2$ with respect to all considered epistemic values $1, ..., n, ..., t$, then $T_1$ should be preferred to $T_2$ overall. Since this Pareto condition is likely not satisfied, this constraint does not result in a complete order of the alternatives under consideration (Bradley 2016, p. 9-10). Within these partial orders, scientists make choices based on their subjective interpretation of values and their endorsing of particular trade-offs

between the values (Bradley 2016, p. 17-18).

I have one worry regarding Bradley's proposal. As he himself points out (see Bradley 2016, p. 10-11) one might ask how much the epistemic values constrain theory choice. The weak Pareto condition might not be satisfied and, hence, all the alternatives are on the table despite the fact, for example, that a theory is just a little bit simpler than another that is empirically much more accurate. Bradley suggests that even in this case there is a lot of constraint given one takes into account that scientists have preferences over alternatives. Having these preferences requires that these preferences satisfy the properties needed for a representation theorem to apply (Bradley 2016, p. 16-17). In my view, these further constraints do not address the question that motivates the debate about theory choice. One is interested in whether the intersubjective, epistemic values constrain the choice between alternatives; not whether additional facts about individual scientists perform this function.

### The universal domain condition

Morreau (2014, 2015) argues that the postulate of unrestricted domain (**U**) should not be applied to theory choice. He argues that some epistemic values can only deliver one profile of alternatives, and, hence, to require that a permissible aggregation function yields a transitive and complete all-things-considered ranking for all possible combinations of profiles is beside the point. In particular, he claims that simplicity can only provide one profile of the alternatives for metaphysical reasons: the simplicity ranking is determined by the way the world is. To put it differently, if a theory is simpler than another this could not have been otherwise (Morreau 2014, p. 9).[9] He goes on to claim that if unrestricted domain does not apply to the situation of theory choice, then the threat of impossibility is avoided (Morreau 2014, p. 10).

Okasha (2015) offers a careful discussion of Morreau's claims. I agree with his general conclusion that although Morreau puts forward a convincing case

[9]Contrast this with the case of empirical accuracy. If the data had been different, another theory might be more accurate than the one previously described as most accurate (Okasha 2015, p. 12).

for some domain restrictions to be relevant in the case of theory choice, these restrictions are not sufficient to rule out the Arrovian impossibility result. The assumption of universal domain is too strong to derive the Arrovian impossibility result. To put it differently, even if universal domain is violated, the space of profiles can still be such that there is no aggregation function that satisfies the other Arrovian conditions (Okasha 2015, p. 11).

### The independence of irrelevant alternatives condition

There are multiple contributions that suggest enriching the informational basis to avoid the Arrovian impossibility result. This amounts to giving up the independence of irrelevant alternative condition (**I**). To recap, this condition requires that the overall ranking of any $T_1$ and $T_2$ depends only on how the epistemic values rank $T_1$ and $T_2$ and not on how they rank other theories in relation to $T_1$ and $T_2$.[10]

Okasha himself sees this as the most promising strategy to avoid the impossibility result. Note that in Arrow's framework, the information that is contained in the rankings of individuals is fairly limited: First, it is assumed that the rankings are purely ordinal, hence, they do not reveal anything about the intensity of preferences; second, these rankings are interpersonally not comparable. Transferring these two points to the case of theory choice amounts to claiming that the epistemic values only provide ordinal rankings and that these rankings cannot be compared across different values (Okasha 2011, p. 98). Okasha suggests paying more attention to what kind of information is provided by the epistemic values. In the simplest form, he suggests scoring the alternative theories under considerations on a common cardinal scale (Okasha 2011, p. 101). Although he argues that in the large-scale theory comparisons that Kuhn had in mind (e.g., the choice between the geocentric and heliocentric theory of the solar system) ordinal information is all that we have, he suggests that in more local theory choices (especially in cases where

[10]The independence of irrelevant alternatives condition (**I**) can be decomposed into two sub-conditions: the independence of irrelevant utilities (IIU) and the ordinal non-comparability of utility (ONC). The conjunction of (IIU) and (ONC) is logically equivalent to (**I**) (Okasha 2015, p. 3).

statistical information about empirical accuracy is available) this enrichment move might be defensible (Okasha 2011, p. 102-103).

Stegenga (2015) points out that the comparability requirements that are needed to avoid the Arrovian impossibility result in the case of theory choice are only rarely met. Okasha agrees with this assessment and, therefore, sees his proposal only as a partially successful escape route (Okasha 2015, p. 5). In a similar spirit, Bradley (2016, p. 13) remarks that it is not clear whether a universal and objective measure for scientific virtues can be defined.

An alternative route to giving up the independence of the irrelevant alternatives condition is suggested by Rizza (2014). He argues that Arrow's impossibility result disappears if one uses the correct information encoded in the ordinal rankings supplied by voters. In the situation of three alternatives, using sequences of three items instead of triples of pairs of alternatives as input for the aggregation function avoids the potential for an intransitive overall ranking (Rizza 2014, p. 1851). The reason for this is that the use of sequences of triples mathematically rules out the possibility for intransitive all-things-considered rankings (Rizza 2014, p. 1851). This translates to the situation of theory choice and, hence, allows one to avoid Okasha's impossibility result.[11]

### 6.4.3 Locating my reply to Okasha

Against the background of these positions in the literature, I can now state more precisely how my solution strategy fits in. The response I develop rejects the analogy between social choice and theory choice. I argue that the theory choice problem poses itself differently in scientific practice; the problem is best described in a cardinal context that allows an explicit weighting of the epistemic values. The upshot is that theory choice outcomes are best described as *all-things-considered judgements*. I devote the rest of this chapter to motivate this switch in the problem description and to propose a concrete method of weighting the different epistemic values.

[11] I will not engage with this suggestion here since this point is a purely mathematical one regarding the characterisation of Arrow's original result. Furthermore, Rizza does not, according to my understanding, discuss the situation in which more than three alternatives are considered. Hence, the generality of his escape route is not clear to me.

Before I turn to this, let me clarify my solution strategy further by highlighting its normative status. Is my solution strategy meant to be a descriptively accurate picture of theory choices in the sciences? Or is it a normative proposal stating how theory choices should be done in the sciences?

I do not aim at providing a descriptively accurate account of theory choice processes in the sciences. By this, I mean that scientists do not necessarily form their beliefs and come to their judgements as specified in the aggregation procedure outlined and defended below. In contrast, I understand my procedure as a rational reconstruction of theory choice processes. To be more precise, my procedure belongs in the normative domain to the extent that it can be viewed as one of multiple possible prescriptive procedures to arrive at an overall judgement in theory choice processes, and, hence, as a normative standard to judge outcomes of theory choice processes. Therefore, my proposal should *not* be viewed as normative in the sense that it is the only aggregation procedure that can defensibly be adopted to arrive at rational theory choices. Rather, my proposal is normative in the sense that it can serve as a benchmark to judge theory choice outcomes since it is one of multiple defensible procedures to resolve instances of theory choice.

## 6.5 A new approach: Using scoring functions over qualitative verdicts to establish comparability of theory choice criteria

In this section, I develop my reply strategy to Okasha's impossibility result. I begin by motivating an alternative description of the theory choice situation that shows that preferences over theories are all-things-considered judgements (*Section 6.5.1*). I then provide an analysis of these all-things-considered judgements in a cardinal context. I argue that Gaertner and Xu (2012)'s general scoring rule can illuminate these types of judgements (*Section 6.5.2*). To show that their general scoring rule is applicable to the problem of theory choice, I proceed in a stepwise fashion. First, I introduce the basic idea behind the scoring rule with the help of a decision situation in a committee. Second,

I flesh out the parallel between the decision situation in a committee and a theory choice situation. I pay particular attention to the notion of context-dependency, which is introduced through the general scoring rule. Third, I provide a formal characterisation of the general scoring rule and comment on the relation of the scoring rule to the Arrovian conditions.

### 6.5.1 Re-thinking the description of theory choice: all-things-considered judgements

Let me return to Kuhn's historically well-informed discussion. Kuhn (1977, p. 323) argues that accuracy, as arguably the central epistemic value, is "by itself seldom or never a sufficient criterion for theory choice". This is because competing theories are often accurate in different domains which necessitates a judgement about which of the domains is more relevant. Kuhn gives the following compelling example:

> The oxygen theory (...) was universally acknowledged to account for observed weight relations in chemical reactions, something the phlogiston theory had previously scarcely attempted to do. But the phlogiston theory, unlike its rival, could account for the metals being much more alike than the ores from which they were formed. (Kuhn 1977, p. 323)

This observation motivates his claim that multiple epistemic values are relevant for choosing between competing theories (Kuhn 1977, p. 321-322). Now, with respect to this set of relevant epistemic values, Kuhn claims that scientists engage in some sort of *weighting procedure*: He claims that scientists can differ about "the relative weights to be accorded to (...) [the] criteria when several are deployed together" (Kuhn 1977, p. 324) and considers the possibility of an "appropriate weight function (...) for their joint application" (Kuhn 1977, p. 326). Kuhn gives the following reason for the necessity of such weighting considerations:

> (...) when [the epistemic values] are deployed together, they repeatedly prove to conflict with one another; accuracy may, for

example, dictate the choice of one theory, scope the choice of its competitor. (Kuhn 1977, p. 324)

These observations suggest that one should view preferences over scientific theories as all-things-considered judgements. The all-things-considered clause refers mainly to the epistemic values that influence the choice of scientists.[12] The prominent role of these all-things-considered judgements in these weighting considerations already goes some way towards undermining Okasha's claim that there is an analogy between social choice and theory choice. For the examples discussed by Kuhn suggest that consistency requirements, that is, the consistency of aggregation rules with desiderata in an ordinal social choice framework, do not appear as relevant input factors for making theory choice decisions. However, to show that Okasha's picture is indeed misguided in important respects, I think it is necessary to sketch a compelling alternative of theory choice. I turn to this task now.

### 6.5.2 Fleshing out the notion of all-things-considered judgements

Talk about preferences over theories as weighted all-things-considered judgements needs to be made more precise. How could one arrive at these all-things-considered judgements in a rationally justifiable way? Can one give a particular procedure to achieve this? If so, what is the relation of this procedure to the Arrovian conditions? I address these questions in order.

#### Starting point: Decision making in a committee

Imagine that you are one of the members of a committee that must decide between a number of research proposals for funding. Suppose that $k$ proposals were submitted. Assume further that the chairperson of your committee comes forward with the following procedure. She declares that there are $m$ categories (from excellent to fail with $m-2$ categories in between), with

[12] Kuhn (1977) also makes clear that subjective factors related to a particular scientist influence the choice. I return to the role of subjective factors in Section 6.6.

rank scores from $m$ to 1 attached to these categories. The chairperson asks all members of your committee to allocate the $k$ proposals to the $m$ available categories. It is not required that every member comes up with a strict ordering and that all categories must be filled by each committee member. Furthermore, the chairperson announces that, as soon as each member has assigned the $k$ proposals to the $m$ categories, she will count the rank numbers assigned to each proposal and then construct a ranking over the $k$ proposals from the highest rank sum to the lowest. The proposal with the highest aggregate rank sum is declared as the winner. More than one proposal may be selected depending on the available budget. The procedure outlined in this example is Gaertner and Xu (2012)'s general scoring rule.[13]

This aggregation procedure can be made fruitful for the case of theory choice. Let me flesh out this analogy in more detail.

### The analogy between the committee's choice and theory choice

In the above scenario, replace the research proposals with alternative theories, the members of the committee with Kuhn's epistemic values, and the chairperson with an individual scientist. Furthermore, consider a set of discrete verdicts corresponding to the categories "very high", "high", "satisfactory", "just sufficient", and "insufficient".[14] This yields the following set-up for theory choice: A scientist considers the alternative theories through the lenses of the epistemic values. For each epistemic value, she independently assigns a qualitative verdict (e.g., "$T_1$ is just sufficiently accurate", "$T_2$'s accuracy is very high") to the alternative theories. The five qualitative verdicts constitute a discrete scale with rank scores. The overall ranking of the theories is determined by the sum of rank scores of each alternative theory.

Drawing this parallel between the case of the committee and the problem

[13]Let me highlight the fact that Gaertner and Xu (2012)'s general scoring rule is not identical to the Borda rule, since it does not presuppose a strict ordering of the alternatives. I will say more about this when I present the formal characterisation of this general scoring rule below.

[14]I will say more about the details of this scale. For the moment, the reader should not be irritated by the fact that I specify an exact number of grades and choose a particular formulation of the qualitative verdicts.

of theory choice amounts to switching to a *cardinal description* of the theory choice problem. Hence, I no longer solve Okasha's challenge in the ordinal context that he sets out. The key question therefore becomes: Why is one justified in introducing this cardinal representation of the theory choice problem? In particular, how can one motivate the qualitative verdicts and their expressions in rank scores?

The presence of a cardinal scale is, in my view, an accurate description of theory choice situations for the following reasons. Theory choice does not take place in a vacuum. Rather, scientists are evaluating competing theories with respect to broad epistemic projects they are conducting in a discipline or a sub-discipline. For example, in engineering such a broad epistemic project could consist in building more efficient combustion engines. In molecular biology, a broad epistemic project could consist in synthesizing new functional germs. In astrophysics, a broad epistemic project could seeking to understand the distribution of clusters of galaxies across the universe. These epistemic projects define a particular assessment context with respect to which the evaluation of competing theories takes place. Looking more closely at these assessment contexts reveals an additional informational structure. To be more precise, evaluating a theory in light of the epistemic values *in an assessment context* allows, for example, to introduce statements about whether a theory is just sufficiently simple, accurate, consistent, fruitful or broad in scope to contribute to the realisation of a particular epistemic project. I assume that the assessment context provides enough informational structure beyond the ordinal information encoded in the epistemic values to assign a sufficiently fine-grained set of qualitative verdicts. Crucially, I do not presuppose that the epistemic values directly provide cardinal information. Rather, I assume an additional step by the scientist which involves careful consideration of the assessment context of a theory choice problem.[15] This is the crucial step in my

[15]Note that the informational content of the assessment context (i.e., the broad epistemic projects and its features) need to be distinct from the epistemic values. If this informational content were identical to the Kuhnian epistemic values or could be rephrased in terms of additional epistemic values, then, given my problem set-up, it would only provide ordinal information and, hence, I would not be able to justify the move to a cardinal scale in this way.

analysis of the notions of all-things-considered preferences over theories: the appropriate language of description of theory choice problems is the cardinal language. However, to justify the use of this cardinal language, one needs to see how the information from the epistemic values (ordinal information) is combined with information about the assessment context of a theory to arrive at cardinal statements in the form of rank scores.

To further motivate the plausibility of the move from ordinal information encoded in the criteria of theory evaluation to a cardinal scale, reconsider the case of the committee. The members of the committee also transfer their ordinal assessment of proposals into a cardinal scale. Following the previous line of reasoning, they can do this by implicitly making assumptions about the research setting at the university or in the discipline more broadly. Accordingly, they are judging whether a research proposal is sufficiently well-structured or original in view of the expectations of the profession, such as the prospect of producing work that could be published in a peer-reviewed journal. These considerations allow the committee members to ascribe qualitative verdicts and ultimately rank scores which then can be aggregated into overall scores for each proposal.

Notice the work that the qualitative verdicts are doing. The qualitative verdicts stated in the rank scoring system impose a cardinal representation on the preference orderings over the alternative theories.[16] It is this constructed cardinal representation that allows inter-criteria comparison. In order to make the cardinal scores for each of the theories comparable across the set of criteria and thereby to achieve inter-criteria comparability, the process of construction of the scale is of importance. To be fully transparent at this stage, my procedure requires establishing a common language amongst the criteria of evaluation.[17]

In order to clarify this construction of a common language, let me go back

[16]See Pivato (2014, p. 50) for a similar discussion of the possibility of imposing cardinality onto a ranking of alternatives.

[17]Note that the idea to introduce qualitative verdicts and thereby establish a common language or "grammar" was made by Balinski and Laraki (2007, 2010). Their proposal of preference aggregation, called majority judgement, however, remains completely within the framework of ordinal information and therefore systematically differs from my own approach.

to the committee and its members again. Each individual must transform their ordinal preference relation over the alternative proposals into a cardinal ranking with the requirement that if proposal $x$ is at least as good as proposal $y$, the cardinal rank or score attached to $x$ is at least as high as the rank assigned to $y$ so that for all $x, y \in X$, the following relationship holds: $xRy \Leftrightarrow s(x) \geq s(y)$, where $s(x)$ stands for the cardinal value or score attached to $x$, and likewise for $y$. This is a very basic requirement in the sense that one must neither lose nor distort ordinal information when one makes a transition from the ordinal to the cardinal world. Furthermore, the cardinality of the chosen ranking system implies that score differences among the different alternatives are meaningful and comparable, so that for four alternatives, $x$, $y$, $z$, $w$, one may come to the conclusion that $s_n(x) - s_n(y) > s_n(z) - s_n(w)$, where $s_n(x)$, for example, is the score assigned to alternative $x$ by committee member $n$. Note that any affine transformation of these scores with a common positive scale factor over all $n$ does not destroy this comparison of score differences.

Coming back to my problem of theory comparison, each scientist is assumed to examine the given theories in the light of the set of criteria that are relevant for the problem at stake. More precisely, each scientist starts for each single criterion with an ordinal ranking over the theories to be evaluated and then transforms this ranking into a sequence of cardinal scores according to the relationship specified above. I assume that the scientist can translate the ordinal into the cardinal information for every epistemic value in isolation. Accordingly, assigning an alternative $x$ a rank score with respect to simplicity, for example, is independent from assigning $x$ a rank score with respect to accuracy, for example. This involves the assumption that the epistemic values are independent of each other. I think that this is, first, in line with Kuhn's discussion and, second, even if there are (conceptual or empirical) dependency relations between the epistemic values, these relations might not hold under all possible interpretations of these values (that should be considered when one discusses theory choice on this level of abstraction).

Since the qualitative verdicts establish the inter-criteria comparability, let me motivate them further. First, the success of my solution does not depend on the particular formulation of qualitative verdicts. I can allow for a

more fine-grained or less fine-grained set of qualitative verdicts. In addition, different formulations of the qualitative verdicts could be chosen for different areas in science. However, the following requirements should be fulfilled by a plausible set of qualitative verdicts: a) the qualitative verdicts need to be framed in evaluative terms. The evaluative terms transport a substantial meaning that can be made sense of in the context of theory choice (e.g., a highly fruitful theory). Furthermore, using evaluative terms instead of going directly for the rank scores implies the commitment to justify the ascription of a particular evaluative term; b) the evaluative terms need to suggest a natural ranking amongst them; c) the evaluative terms need to make sense with respect to every epistemic value under consideration.

Second, let me provide some plausibility for the claim that my five qualitative verdicts "very high", "high", "satisfactory", "just sufficient", and "insufficient" are indeed applicable to the five Kuhnian epistemic values. In my view, it is fairly obvious that accuracy, scope, fruitfulness and simplicity can be fulfilled to a greater or lesser extent. It is less obvious for consistency. Under the heading of this epistemic value, Kuhn discusses internal consistency (i.e., that a theory is free of any contradictions) and external consistency (i.e., that a theory does not contradict already accepted theories) (Kuhn 1977, p. 321). I accommodate internal and external consistency in my set of qualitative verdicts as follows: If the theory contains no contradictions and does not entail contradictions with already existing theories, it receives the verdict "very high", otherwise it receives the verdict "insufficient". Accordingly, I treat "consistency" as a matter of binary choice. Importantly, I do not treat the qualitative verdict "insufficient" as an eliminative verdict in the sense that whenever with respect to one epistemic value an alternative receives an "insufficient" verdict, this alternative is eliminated from the choice set. The reason for this is that I, in line with Kuhn, do not share the intuition that one of the epistemic values should be treated as a "killer" criterion. Accordingly, even if a theory is insufficiently simple, let me say, but receives the best qualitative verdicts with respect to all other criteria such that the aggregate rank score is the highest of all alternatives, the low grade in terms of simplicity should be seen in relation to the high grades obtained from the other criteria.

Third, the talk about theory choice problems in an assessment context introduces a notion of context-dependency. Let me be clear what dependency relations I have in mind here. In my framework, the graininess of the partition of the set of qualitative verdicts can vary between assessment contexts. Furthermore, the prerequisites in order to assign the qualitative verdict of "sufficient", for example, to a particular theory can vary among the epistemic values and may depend on the assessment context. What is important, however, is that if the scientist has come to the conclusion that two criteria are sufficiently fulfilled with respect to any particular theory or across two or more theories, then, simply said, "sufficient means sufficient". Otherwise, the inter-criterion comparability would not be given.

**More on the notion of context-dependency**

In this section, I provide further motivation for my proposal of shared qualitative verdicts in the context of theory evaluation. So far, I have argued that the theory choice problem should be viewed in a cardinal context and that qualitative verdicts can be imposed on all of the Kuhnian values. Now I show that the set of qualitative verdicts allows for considerable flexibility by spelling out the element of context-dependency. I do this in two steps. First, I motivate the claim that the level of graininess of the qualitative verdicts can vary among assessment contexts. Second, I argue that the prerequisites in order to reach a particular qualitative verdict can change across epistemic values and assessment contexts.

To start, think about the following two hypothetical examples of assessment contexts. Martina, a particle physicist at CERN, the European Organisation for Nuclear Research, evaluates two theories about the structure of the decay of Higg's Bosons. Tom, a sociologist, evaluates two theories about the causes of the recent increase in immigration to the United Kingdom. Martina and Tom are using the five Kuhnian values to reach an overall ranking of theories. They impose a set of qualitative verdicts on their ordinal preferences for each criterion. To do this, Martina and Tom might be using different sets of qualitative verdicts. Martina might work, in light of the small differences in

the content of the theories and the necessary precision of the predictive tasks, with a five-item scale. In contrast, Tom might be using a three-item scale that is appropriate for dealing with the recent aggregate data on immigration flows.

Let me turn to the prerequisites to reach a particular qualitative verdict in the same assessment context. Assume that Martina is using the following verdicts: "very high", "high", "sufficient", "just sufficient", and "insufficient". When she assigns the competing theories to the qualitative verdicts, she reviews the ordinal information provided by every epistemic value. Now, to do this pairing of theories and verdicts it is, as I asserted earlier, it is absolutely necessary that a qualitative verdict (e.g., "insufficient") means the same for every epistemic value. The prerequisites to reach a particular qualitative verdict can, however, be quite different. With respect to accuracy, for example, "high" could refer to a specific number of decimals at which the prediction of a theory matches the data. With respect to simplicity, "high" could denote the fact that a theory allows stating the key differential equation for the system under study in closed form. The same could be argued for the other qualitative verdicts.

What about the prerequisites to reach a particular verdict in different assessment contexts? Assume that Tom and Martina are using the same set of verdicts, as specified in the previous paragraph. Furthermore, assume that both of them attach the verdict "high" to one of their theories with respect to accuracy. Since their application contexts (particle physics vs. sociology) differ substantially, the reasoning behind the respective verdicts can differ. Martina could interpret accuracy as a specific number of decimals at which the prediction of a theory matches the data. In contrast, Tom could refer to the fact that one of his theories is able to reflect qualitatively what people have reported in narrative interviews.

### A formal characterisation of the general scoring rule

Let me conclude my attempt to make the notion of all-things-considered judgements more precise in a cardinal context, by providing a formal charac-

terisation of Gaertner and Xu (2012)'s general scoring rule. Let $X$ be the set of scientific theories containing a finite number of elements. Let $N$ be the set of criteria deemed relevant with $t > 1$. Let $E = 1, ..., E$, with the cardinality of this set being larger than one, be a set of given positive integers from 1 to $E$. These integers will in most cases be assumed to be equally distanced and are thought to represent qualitative statements thus constituting a common language of evaluation, as outlined above.[18]

A scoring function $s_i$: $X \to E$ is chosen for each criterion $i \in N$, such that, for all $x \in X$, $s_i(x)$ indicates the score that criterion $i$ assigns to $x$. Let $S_i$ be the set of all possible scoring functions for criterion $i$. As explained in the last section, the statement how well or how badly a theory fares in light of a criterion must be inserted in the commonly given scale constituted by set $E$.

Let $P$ be the set of all orderings over $X$. A profile $s = (s_1, ..., s_t)$ is a list of scoring functions, one for each criterion. An aggregation rule $f$ is defined as a mapping: $S_1 x ... x S_t \to P$. Let $S = S_1 x ... x S_t$.

$f$ is said to be an $E$-based scoring rule, to be denoted by $f_E$, if and only if, for any $s \in S$, and any $x, y \in X$, it is the case that:

$$x \succeq y \Leftrightarrow \sum_{i \in n} s_i(x) \geq \sum_{i \in n} s_i(y)$$

where $\succeq = f(s)$. The asymmetric and symmetric parts of $\succeq$ will be denoted by $\succ$ and $\sim$, respectively.[19]

I am now in the position to address the question as to how my reply strategy relates to the Arrovian conditions (**U**, **P**, **D**, and **I**).

The general scoring rule does not restrict the set of possible binary relations $R_n$ over the alternative theories. Hence, unrestricted domain (**U**) is satisfied. The weak Pareto condition (**P**) is also satisfied. If all epistemic

[18] In general, the required minimal level of graininess depends on the particular theory choice problem at hand. However, if one criterion ($i$) ranks all alternative theories in a strict order, then due to $xR_iy \Leftrightarrow s_i(x) \geq s_i(y)$, the minimal level of graininess is the number of alternative theories.

[19] This $E$-based scoring function can be characterised axiomatically in a fairly simple way. See Gaertner and Xu (2012) for a formal characterisation.

values find $T_1$, for example, better than $T_2$, then according to the assumption that $xRy \Leftrightarrow s(x) \geq s(y)$ and $xPy \Leftrightarrow s(x) > s(y)$, respectively, the Pareto condition requires that $T_1$ is better than $T_2$ in the "world" of cardinal information. Furthermore, the non-dictatorship (**D**) requirement states that there is no epistemic value such that if it ranks $T_1$ above $T_2$, $T_1$ is ranked automatically above $T_2$ in the overall ranking. This is satisfied by the general scoring rule because the general scoring rule weights the rank scores of each criterion equally.[20] Finally, the independence of irrelevant alternatives (**I**) requirement means that the overall ranking of $T_1$ and $T_2$ depends only on how the epistemic values rank $T_1$ and $T_2$ and not on how they rank other alternatives. The proposed aggregation procedure satisfies condition **I** reformulated within the cardinal context.[21] Verbally, it requires that if two theories $T_1$ and $T_2$ receive precisely the same rank scores from the different epistemic values in the case of two separate evaluations, then the aggregate judgement over $T_1$ and $T_2$ is identical between the two evaluations. Given this set-up, it is irrelevant for the aggregate ranking of $T_1$ and $T_2$ how other theories are evaluated in the two evaluations. Let me illustrate the cardinal version of the independence condition with the help of a simple example. Consider the two evaluations of three theories $T_1$, $T_2$ and $T_3$ based on three epistemic values where the ranks or scores in the left column are embedded in a simple integer-valued, equally-spaced interval scale (see Tables 6.2 and 6.3).

According to the general scoring function, in the first evaluation $T_1$ (with an associated total rank score 6) is strictly preferred to $T_2$ (having an associated total rank score 5). These exact scores with respect to $T_1$ and $T_2$ are retained in the second evaluation, and, therefore, the aggregate relations between the two theories are exactly the same despite the fact that an irrelevant alternative (here $T_3$) is positioned differently in the two evaluations.

[20] The scoring rule allows attaching different weights to the epistemic values. One way to account for this difference is to divide a criterion (e.g., "fruitfulness") into two subcriteria (e.g., "fruitfulness with respect to the discipline" and "fruitfulness with respect to neighbouring disciplines"). In this way, the initial criterion gets a higher weight in the summation procedure. Ascribing different weights might be in order if one recognises different types of values within Kuhn's epistemic values as Douglas (2013) suggests.

[21] A discussion of the reformulated version of the independence requirement in the context of the utilitarian rule can be found in Gaertner (2013, p. 125-126).

| | | | |
|---|---|---|---|
| 4 | | | |
| 3 | $T_1$ | $T_3$ | $T_3$ |
| 2 | $T_2$ | $T_1$ | $T_2$ |
| 1 | $T_3$ | $T_2$ | $T_1$ |
| 0 | | | |

Table 6.2: Evaluation 1

| | | | |
|---|---|---|---|
| 4 | $\mathbf{T_3}$ | $\mathbf{T_3}$ | |
| 3 | $T_1$ | | |
| 2 | $T_2$ | $T_1$ | $T_2$ |
| 1 | | $T_2$ | $T_1$ |
| 0 | | | $\mathbf{T_3}$ |

Table 6.3: Evaluation 2

Finally, it seems to me that the method proposed is not only distinct from but also superior to the Borda rule. While the latter rule requires that each and every criterion ranks the alternative theories in a linear order, such a high degree of uniformity is not demanded by the method proposed here. Different criteria can rank or rather assign scores to the given alternatives in completely different ways as explained previously. I consider this as an advantage since the single criterion has more flexibility to express to what degree or extent it finds itself represented among the various theories under consideration.

## 6.6 Aggregation across different scientists

What I have described in the last section essentially is an aggregation procedure that allows resolving theory choice problems of a *single* scientist. Different scientists will normally come up with different orderings over the theories to be evaluated. Kuhn (1977, p. 325) writes that "every individual choice between competing theories depends on a mixture of objective and subjective factors, or of shared and individual criteria". He goes on to say:

> (...) I have conceded that each individual has an algorithm and that all their algorithms have much in common. Nevertheless,

> I continue to hold that the algorithms of individuals are all ultimately different by virtue of the *subjective* considerations with which each must complete the objective criteria before any computations can be done. (Kuhn 1977, p. 329, my emphasis)

According to my proposal, each and every scientist can rely on his or her own scoring function in order to generate an ordering over alternative theories. Additionally, different scientists could choose different degrees of graininess with respect to qualitative verdicts. One person could rely on three verdicts, for example "high", "sufficient" and "insufficient", another person could decide on only two verdicts, say "high" and "insufficient". If this is the case, the two scientists generate different rankings of the theories, as in the following example.

Consider Anna and Peter, two scientists who evaluate two theories $T_1$ and $T_2$ with two different sets of qualitative verdicts (see Tables 6.4 and 6.5).

| Verdicts | Rank scores | Criterion 1 | Criterion 2 | Criterion 3 |
|---|---|---|---|---|
| high | 3 | $T_1$ | | |
| sufficient | 2 | | $T_2$ | $T_2$ |
| insufficient | 1 | $T_2$ | $T_1$ | $T_1$ |

Table 6.4: Scientist Anna's evaluation of two competing theories $T_1$ and $T_2$.

| Verdicts | Rank scores | Criterion 1 | Criterion 2 | Criterion 3 |
|---|---|---|---|---|
| high | 2 | $T_1$ | $T_2$ | $T_2$ |
| insufficient | 1 | $T_2$ | $T_1$ | $T_1$ |

Table 6.5: Scientist Peter's evaluation of two competing theories $T_1$ and $T_2$.

Notice that in terms of purely ordinal information, Anna and Peter reveal the same preference ordering. Using the scoring procedure, Peter strictly prefers $T_2$ over $T_1$ whereas Anna is indifferent between $T_1$ and $T_2$.

As can be seen, Anna and Peter assign $T_1$ and $T_2$ to different qualitative verdicts. The reasons for their disagreement in the assignment of qualitative verdicts might be that Anna and Peter interpret the qualitative verdicts differently and/or that they interpret the epistemic values in a different manner.

All this can happen. One person may never assign the grade "very good" – because for this person "good" is the best ever. Another person may be easily satisfied and, therefore, assign the grade "very good" quite often. As I stated before, once the grades have been assigned, they must be taken at face value. One should assert that the grade "good" of one person is equivalent to the grade "very good" of another person. One simply lacks this information. Such statements could become acceptable only under special circumstances where one has detailed information about the personality and psychology of different persons, which normally is not the case.

Nevertheless, once an overall verdict among a group of different scientists is found necessary, there is need for a mechanism that aggregates across individual evaluations. The procedure put forward in this chapter establishes, for each evaluating person, an ordinal ranking over the set of alternative theories at stake. Various methods are available to aggregate these orderings, all of which violate at least one of Arrow's requirements. The Borda rule is one candidate, approval voting (Brams and Fishburn 1983) would be a second, plurality voting a third, Condorcet's pairwise majority voting rule a fourth. I shall now offer some considerations for why the Borda rule might be an appropriate procedure for aggregation *across* different scientists.

To start, notice that it does not make sense to use Gaertner and Xu (2012)'s general scoring rule at this stage. To do so would be asking the scientists to use a set of qualitative verdicts to transform their ordinal rankings into cardinal information. However, this would amount to saying that the scientists need to go back to their individual assessment of the alternative theories instead of taking the outcome of this analysis at face value. Furthermore, the Borda method has an advantage in relation to plurality and approval voting, namely, its aggregation procedure uses a lot of positional information that both plurality and approval voting ignore. While the plurality rule restricts itself to using information on the top element within each person's evaluation only so that the ranking of all other options is ignored, approval voting implicitly constructs two indifference classes, the set of acceptable options and the set of unacceptable alternatives, with no further differentiation in either set. Finally, the Borda method might be preferred

to Condorcet's pairwise majority voting since the former guarantees that one never receives a cyclical ranking of the alternative theories on the level of the scientific community.

## 6.7 Concluding remarks

In this chapter, I have engaged with Okasha's impossibility result regarding multi-criteria theory choice to address my fifth research question: What should one do if one faces a situation of non-robust evidence from multiple evidential sources? I showed that one is not always in the epistemic position where one can suspend one's judgement. I have argued that such situations can be characterised by four features: a) alternative options have radical different outcomes given different states of the world, b) one cannot suspend judgement due to the time sensitivity of the choices, c) there is sparse information about the decision-relevant states of the world, which prohibits (or at least makes it difficult) to use standard decision theoretic tools, and d) the evidence, which could help resolve the factual issue about what the relevant future state of the world is for the decision maker, consists of or is underlied by models or theories. In these situations, theory or model choice is a sensible step to take.

I have put forward a procedure for making such theory or model choices. I have argued that preferences over theories are best described as weighted all-things-considered judgements. These all-things-considered judgements can be analysed in a cardinal context. Moving to a cardinal context allowed me to rely on Gaertner and Xu (2012)'s general scoring rule to put forward a procedure for making theory choices. Finally, I have argued that my solution can capture Kuhn's statements about the role of subjective factors in the theory choice process. I have suggested the Borda rule as a suitable method of aggregating the rankings across different scientists.

# Chapter 7

# Concluding Remarks

This thesis started with two episodes from medical and social policy-making. I highlighted the fact that, until the late 19th century, medicine did not provide effective treatments for many diseases. Something similar held true for interventions in the housing market of late Victorian United Kingdom, which wrestled unsuccessfully with the consequences of rapid industrialisation. To be effective, policy interventions should be based on the best available evidence about the domain in which one is intervening. Evidence-based policy-making addresses the task of improving the effectiveness of policy interventions via evidence.

One of the key issues that this approach to policy-making faces is that a large amount of potentially conflicting evidence from different sources can be relevant to a particular problem. Robustness analysis emerged as a tool to deal with this situation of evidential diversity. I have strengthened the case for the use of robustness analysis in evidence-based policy-making by answering open research questions on this inference technique. The starting point consisted of a review of the current state of the philosophical literature on robustness analysis. The review revealed the need to address the following five questions about robustness analysis:

1. Do measurement and derivational robustness analysis exhaust the set of useful types of robustness analysis?

2. What is the value of derivational robustness analysis if not the entire

relevant model space is covered by the available models?

3. What should one do if the model results are non-robust, that is, the model derivations are not in agreement?

4. How should one conceptualise the relation between expert knowledge and the evidence basis when one applies the framework of measurement robustness?

5. What should one do if one faces a situation of non-robust evidence from multiple evidential sources?

In response to these questions, I have articulated and defended five claims. I now summarise these claims and locate my contribution within the philosophical debate on robustness analysis.

Based on an analysis of toy models from a range of scientific disciplines, I argued that the stability of model results across different (types of) target systems – what I called *predictive stability* – is a further category of robustness considerations. I provided a formal inference scheme for this type of robustness analysis. I showed that this inference scheme differs from derivational robustness analysis by being directly about the relation of model results and target systems as well as by not requiring a change in the model assumptions. Furthermore, I argued that predictive stability is distinct from a key tenet of confirmation theory, which holds that diverse evidence better confirms a hypothesis than uniform evidence, because it solely postulates the presence of a model result in target systems and does not in itself raise the issue of confirmation. Taken together, I think this substantiates the need to expand the taxonomy of robustness analysis provided by Woodward (2006).

With respect to the second research question, I argued that derivational robustness analysis has value even if less than the entire relevant model space is covered. Its value consists in the fact that it provides the resources to assess in a structured way how much of the relevant possibility space is covered by a model. As my case study of Hunter and Williams's automated evidence aggregator showed, this fact can be used to formulate preference relations over automated evidence aggregation procedures and, hence, can serve as an

evaluation criterion of automated evidence aggregators. My analysis suggests re-thinking the point of application of derivational robustness analysis. In contrast to seeing it as an ex post tool for checking the stability of model results, as has been the focus of the literature on the confirmatory import of this inference technique, derivational robustness analysis can be employed *ex ante* in the process of model construction or, as shown in the case study, of algorithm design.

I went on to argue that, with respect to the third research question, even in cases of a lack of derivational robustness models can provide a form of conceptual learning or introduce a potentially valuable modelling technique. I drew this conclusion based on a distinction between learning from and learning with a toy model as well as on a detailed analysis of Hotelling's model of minimal differentiation. The conceptual learning from this non-robust toy model consists in the fact that one sees what different sets of assumptions imply, and, hence, one ends up with a better grasp of a conceptual scheme. The point about the modelling technique consists in the claim that toy models can show how a modelling technique can be applied fruitfully to a new question or a new domain. I argued that both of these insights are only of limited use for economic policy-making, since they do not provide the relevant knowledge about causal relationships that licences policy-interventions. Hence, I urged economic policy-makers to be more transparent about the epistemic goals that they seek to achieve with such toy models.

With respect to the fourth research question, I argued that expert knowledge can only under special circumstances be viewed as a separate evidential mode that stands on a par with other evidential modes, such as observations or a body of theoretical knowledge, in measurement robustness considerations. My engagement with the IPCC uncertainty framework showed that expert knowledge – in the form of individual expert judgement or collective expert agreement – is a necessary type of knowledge for addressing questions that arise when one applies the measurement robustness framework; questions such as whether evidential modes are independent and whether they are of high quality. My two claims make the relation between expert knowledge and measurement robustness analysis transparent and, thereby, open it up

for further scrutiny. A spin-off from the discussion of the IPCC framework is my suggestion to improve the framework, that is, to give up the bifurcation between evidence and expert agreement, to provide further categories for the assessment of the available evidence, and to employ tools from social choice theory and multi-criteria decision analysis to aggregate these categories.

Finally, I argued that in specific decision situations in which an epistemic agent faces conflicting evidence from different evidential sources, selecting a theory or a model, which underlies or constitutes a subset of the conflicting evidence, is a reasonable step to take and I provided a procedure to make this choice in a rationally defensible way. The procedure suggests a way of seeing theory or model choices in a cardinal context. Once this cardinal context is established, one can apply scoring to make the choice. Scoring rules allow the aggregation of information from multiple criteria that are deemed relevant for the choice problem. My proposal rejects Okasha's analogy between theory choice and social choice and shows that theory choice in a scientific context is best accounted for in a weighting and scoring framework.

Taking a step back from the details of these five claims and the case studies that back them up, reveals two central insights about robustness analysis as a tool for evidence-based policy-making. First, the facts that hypotheses are robust model results or are supported by a varied set of evidence bear important epistemic weight in the context of evidence-based policy-making. The conditions that need to be in place such that derivational and measurement robustness confirm hypotheses can be met in policy domains. My answers to research questions one and four have complemented the already existing arguments for the epistemic import of robustness by providing a new set of relevant variations, that is, variation across different types of target systems, and by showing how expert knowledge should be combined with evidential claims. Second, however, my analysis points to the fact that when robustness reasoning is applied to policy questions, often the conditions for the confirmatory import of robustness are not met: not the entire relevant model space is covered, the model results are not robust, or different evidential modes are not in agreement. I characterised and motivated such situations along my research questions two, three, and five. I argued that even in these three

cases, the inference technique of robustness analysis yields fruitful insights for evidence appraisal. In particular with respect to this latter insight, my thesis strengthens the case for the use of robustness analysis in evidence-based policy-making.

# Bibliography

Adler, C. E. and Hadorn, G. H. (2014). The IPCC and treatment of uncertainties: Topics and sources of dissensus. *WIRES Climate Change*, 5:663–676.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500.

Akerlof, G. A. (2001). Writing the "market for 'lemons'": A personal and interpretativ essay. *The Sveriges Riksbank Prize in Economic Science in Memory of Alfred Nobel 2001*, pages 1–10.

Alexander, J. (2007). *The structural evolution of morality*. Cambridge University Press, Cambridge.

Alexandrova, A. and Northcott, R. (2013). It's just a feeling: why economic models do not explain. *Journal of Economic Methodology*, 20(3):262–267.

Arrow, K. J. (1963). *Social Choice and Individual Values*. John Wiley, New York.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, 463(7279):294–295.

Aven, T. and Renn, O. (2015). An evaluation of the treatment of risk and uncertainties in the IPCC reports on climate change. *Risk Analysis*, 35:701–712.

Aydinonat, E. (2006). An interview with Thomas C. Schelling: Interpretation of game theory and the checkerboard model. *Economics Bulletin*, 2(2):1–7.

Ayer, A. J. (1956). *The problem of knowledge*. St Martin's Press, New York.

Balinski, M. and Laraki, R. (2007). A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences*, 104:8720–8725.

Balinski, M. and Laraki, R. (2010). *Majority Judgment: measuring, ranking and electing*. MIT Press, Cambridge MA.

Battermann, R. W. and Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, 81(3):349–376.

Ben-Gal, I. (2007). Bayesian networks. In Ruggeri, F., Faltin, F., and Kenett, R., editors, *Encyclopedia of Statistics in Quality and Reliability*, pages 1–6. Wiley and Sons.

Beugelsdijk, S., de Grott, H. F., and van Schaik, A. B. (2004). Trust and economic growth: a robustness analysis. *Oxford Economic Papers*, 56:118–134.

Boelhouwer, P. and Hoekstra, J. (2012). Housing and the State in Western Europe. In Smith, S. J., editor, *International Encyclopedia of Housing and Home*, pages 363–373. Elsevier, London.

Bohr, N. (1913). On the constitution of atoms and molecules, part i. *Philosophical Magazine*, 26(151):1–24.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, 180:33–45.

Bokulich, A. (2012). Distinguishing explanatory from nonexplanatory fictions. *Philosophy of Science*, 79(5):725–737.

Bovens, L. and Hartmann, S. (2003). *Bayesian Epistemology*. Oxford University Press, Oxford.

Bradley, R. and Drechsler, M. (2014). Types of uncertainty. *Erkenntnis*, 79(6):1225–1248.

Bradley, R., Helgeson, C., and Hill, B. (2016). Climate change assessments: Confidence, probability and decision. *Philosophy of Science*, pages 1–28.

Bradley, S. (2016). Constraints on rational theory choice. *British Journal for Philosophy of Science*, pages 1–23.

Brams, S. J. and Fishburn, P. C. (1983). *Approval voting*. Birkhäuser, Boston.

Brenner, S. (2001). Determinants of product differentiation: A survey. Technical report, Humboldt University Institute of Management.

Bricmont, J. (1996). Science of Chaos or Chaos in Science. In Gross, P., Levitt, N., and Lewis, M. W., editors, *The flight from science and reason*, pages 131–175. The New York Academy of Sciences, New York.

Briggs, C. J. and Hoopes, M. F. (2004). Stabilizing effects in spatial parasitoid-host and predator-prey models: A review. *Theoretical Population Biology*, 65:299–315.

Brun, G. (2015). Eplication as a method of conceptual re-engineering. *Erkenntnis*, pages 1–30.

Budescu, D. V., Por, H.-H., Broomell, S. B., and Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, pages 1–5.

Bycroft, M. (2009). Going outside the model: Robustness analysis and experimental science. *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 3(1):123–141.

Carnap, R. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research*, 8:133–148.

Carnap, R. (1962). *Logical Foundations of Probability*. The University of Chicago Press, Chicago.

Cartwright, N. (1991). Replicability, reproducibility, and robustness: Comments on Harry Collins. *History of Political Economy*, 23(1):143–155.

Cartwright, N. (1999). *The dappled world*. Cambridge University Press, Cambridge.

Cartwright, N. (2009). If no capacities then no credible worlds. But can models reveal capacities? *Erkenntnis*, 70:45–58.

Cartwright, N. (2010). Models: Parables vs. Fables. In Frigg, R. and Hunters, M., editors, *Beyond Mimesis and Convention*, pages 19–31. Springer, New York.

Cartwright, N. and Hardie, J. (2012). *Evidence-Based Policy – A Practical Guide to Doing It Better*. Oxford University Press, Oxford.

Casini, L. (2014). Not-so-minimal models: Between isolation and imagination. *Philosophy of the Social Sciences*, 44(5):646–672.

Cipra, B. A. (1987). An introduction to the Ising model. *American Mathematical Monthly*, 94(10):937 – 959.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, 33:339–360.

Collins, H. M. (1984). When do scientists prefer to vary their experiments? *Studies in History and Philosophy of Science*, 15(2):169–174.

Contessa, G. (2007). Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science*, 74:48–68.

Cosgrove, L., Vannoy, S., Mintzes, B., and Schaughnessy, A. F. (2016). Under the influence: The interplay among industry, publishing, and drug regulation. *Accountability in Research*, 23(5):257–279.

Cresto, E., del Corral, M., Tajer, D., Nascimbene, J., and Cassini, A. (2017). Confirmational holism and the amalgamation of evidence. In Massimi, M., Romeijn, J.-W., and Schurz, G., editors, *EPSA 2015 Selected Papers. European Studies of Philosophy of Science*, pages 273–284. Springer, Frankfurt.

Culp, S. (1994). Defending robustness: The bacterial mesosome as a test case. *Philosophy of Science*, pages 46–57.

Culp, S. (1995). Objectivity in experimental inquiry: Breaking data-technique circles. *Philosophy of Science*, 62:430–450.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 47:5–20.

de Regt, H. and Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, 144:137–170.

Doob, J. L. (1971). What is a martingale? *American Mathematical Monthly*, 78:451–462.

Douglas, H. (2012). Weighing complex evidence in a democratic society. *Kennedy Institute of Ethics Journal*, 22(2):139–162.

Douglas, H. (2013). The value of cognitive valus. *Philosophy of Science*, 80(5):796–806.

Dragulescu, A. and Yakovenko, V. M. (2000). Statistical mechanics of money. *The European Physical Journal B*, 17:723–729.

Eronen, M. I. (2015). Robustness and reality. *Synthese*, 192:3961–3977.

Fitelson, B. (2001). *Studies in Bayesian Confirmation Theory*. PhD thesis, University of Wisconsin-Madison.

Forber, P. (2010). Confirmation and explaining how possible. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 41:32–40.

Franklin, A. and Howson, C. (1984). Why do scientists prefer to vary their experiments? *Studies in the History of Philosophy of Science*, 15(1):51–62.

Friedli, S. and Velenik, Y. (2016). Statistical mechanics of lattice systems: A concrete mathematical introduction - chapter 1. *Manuscript*.

Frigg, R. (2010). What is statistical mechanics. In Galles, C., Lorenzano, P., Ortiz, E., and Rheinberger, H.-J., editors, *Encyclopedia*, pages 1–27. Eolss, Isle of Man.

Frigg, R. and Hartmann, S. (2012). Models. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*, pages 1–19. Metaphysics Research Lab, Stanford University, spring 2016 edition.

Frigg, R. and Nguyen, J. (2016). The turn of the valve: Representing with material models. *Manuscript*, pages 1–24.

Frigg, R. and Nguyen, J. (2017). Models and representation. In Magnani, L. and Bertolotti, T., editors, *Handbook of Model-Based Science*, pages 49–102. Springer, Dordrecht, Heidelberg, London and New York.

Fuller, J. (2013). Rhetoric and argumentation: how clinical practice guidelines think. *Journal of Evaluation in Clinical Practice*, 19:433–441.

Fumagalli, R. (2015). Why we cannot learn from minimal models. *Erkenntnis*, pages 1–23.

Gaertner, W. (2013). *A Primer in Social Choice Theory*. Oxford University Press, Oxford.

Gaertner, W. and Xu, Y. (2012). A general scoring rule. *Mathematical Social Sciences*, 62:193–196.

Gibbard, A. and Varian, H. R. (1978). Economic models. *The Journal of Philosophy*, 75(11):664–677.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, 71:742–752.

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, 172:269–281.

Gorgoannis, N., Hunter, A., and Williams, M. (2009). An argument-based approach to reasoning with clinical knowledge. *International Journal for Approximate Reasoning*, 51:1–22.

Gottwald, G. A. and Oliver, M. (2009). Boltzmann's dilemma: An introduction to statistical mechanics via the Kac ring. *Society for Industrial and Applied Mathematics*, 51(3):613–635.

Greenstone, G. (2010). The history of bloodletting. *BC Medical Journal*, 52(1):12–14.

Gross, A. M. (2016). *Understanding disease through data driven biology*. PhD thesis, UC San Diego.

Grüne-Yanoff, T. (2009). Learning from minimal economic models. *Erkenntnis*, 70:81–99.

Grüne-Yanoff, T. (2013a). Appraising models non-representationally. *Philosophy of Science*, 80(5):850–861.

Grüne-Yanoff, T. (2013b). Genuineness resolved: a reply to Reiss' purported paradox. *Journal of Economic Methodology*, 20(3):255–261.

Hajar, R. (2012). The air of history (part ii) medicine in the middle ages. *Heart Views*, 13(4):158–162.

Hands, D. W. (2016). Derivational robustness, credible substitute systems and mathematical economic models: The case of stability analysis in Walrasian general equilibrium theory. *European Journal for Philosophy of Science*, 6:31–53.

Hanson, L. P. and Sargent, T. J. (2001). Robust control and model uncertainty. *The American Economic Review*, 91(2):60–66.

Hartmann, S. (1995). Models as a tool for theory construction: Some strategies of preliminary theory construction. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 44:1–20.

Hartmann, S. (1998). Idealization in quantum field theory. In Shanks, N., editor, *Idealization in Contemporary Physics*, pages 99–122. Rodopi, Amsterdam.

Hartmann, S., Reutlinger, A., and Hangleiter, D. (2016). Understanding (with) toy models. *The British Journal for Philosophy of Science*, pages 1–34.

Hausman, D. M. (2013). Paradox postponed. *Journal for Economic Methodology*, 20(3):250–254.

Henderson, L. (2014). Can the second law be compatible with time reversal invariant dynamics. *Studies in History and Philosophy of Modern Physics*, 47:90–98.

Hey, S. P. (2015). Robust and discordant evidence: Methodological lessons from clinical research. *Philosophy of Science*, 82(1):55–75.

Hindriks, F. (2013). Explanation, understanding, and unrealistic models. *Studies in History and Philosophy of Science*, 44:523–531.

Hirsh-Pasek, K., Golinkoff, R. M., Berk, L. E., and Singer, D. G. (2009). *A mandate for playful learning in preschool: Presenting the evidence*. Oxford University Press, New York.

Holling, C. S. (1959). The components of predation as revealed by a study of small-mammal predation of the European Pine Sawfly. *The Canadian Entomologist*, 91(5):234–260.

Hotelling, H. (1929). Stability in competition. *The Economic Journal*, 39(153):41–57.

Howson, C. and Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court, Chicago.

Hoyningen-Huene, P. (1993). *Reconstructing Scientific Revolutions: Thomas S. Kuhn's Philosophy of Science*. University of Chicago Press, Chicago.

Hunter, A. and Williams, M. (2010). Using clinical preferences in argumentation about evidence from clinical trials. *Proceedings of the 1st ACM International Health Information Symposium*, pages 118–127.

Hunter, A. and Williams, M. (2012). Aggregating evidence about the positive and negative effects of treatments. *Artificial Intelligence in Medicine*, 56(3):173–190.

Hunter, A. and Williams, M. (2013). Aggregating evidence about the positive and negative effects of treatments using a computational model of argument. *LSE Choice Group Talk (October 21, 2013)*, pages 1–38.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge.

IPCC (2014a). Intergovernmental Panel on Climate Change: Organization. http://www.ipcc.ch/organization/organization.shtml.

IPCC (2014b). Intergovernmental Panel on Climate Change: Working groups / task force. http://www.ipcc.ch/working_groups/working_groups.shtml.

Ising, E. (1925). Beitrag zur Theorie des Ferromagentismus. *Zeitschrift für Physik*, 31(1):253–258.

Jaeger, D. (2009). Toy models and stylized facts. https://criticalvalue.wordpress.com/2009/06/12/toy-models-and-stylized-facts/.

Jebeile, J. (2016). Learning from a toy model: the Kac ring. *Manuscript*, pages 1–29.

Johnson, P. (2005). Market disciplines in Victorian Britain. *Working Papers on The Nature of Evidence: How Well Do 'Facts' Travel?*, (6):1–32.

Jones, R. N. (2011). The latest iteration of IPCC uncertainty guidance: An author perspective. *Climate Change*, 108:733–743.

Justus, J. (2012). The elusive basis of inferential robustness. *Philosophy of Science*, 79(5):795–807.

Kac, M. (1959). *Probabilities and related topics in physical sciences*. Interscience Publisher, New York.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press, Cambridge.

Kincaid, H. (2001). The ethical and epistemic issues in inferior treatment in clinical research. *Formosan Journal of Medical Humanities*, 2:34–40.

Kincaid, H. (2011). Causal modelling, mechanism, and probability in epidemiology. In Illari, P., Russo, F., and Williamson, J., editors, *Causality in the Sciences*. Oxford University Press, Oxford.

Kincaid, H. and McKitrick, J., editors (2007). *Establishing Medical Reality: Essays in the Metaphysics and Epistemology of Biomedical Science*. Springer, Rotterdam.

Knuuttila, T. and Loettgers, A. (2016). Model templates within and between disciplines: From magnets to gases – and socio-economic systems. *European Journal for Philosophy of Science*, 6:377–400.

Kocherlakota, N. (2009). Federal Reserve Bank of Minneapolis, annual report: Modern macroeconomic models as tools for economic policy. https://www.minneapolisfed.org/publications/the-region/modern-macroeconomic-models-as-tools-for-economic-policy.

Krimsky, S. (2005). The weight of scientific evidence in policy and law. *American Journal of Public Health*, 95(1):129–136.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago, 3rd edition.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In Kuhn, T. S., editor, *The Essential Tension – Selected Studies in Scientific Tradition and Change*, pages 320–339. The University of Chicago Press, Chicago.

Kuorikoski, J. and Lehtinen, A. (2009). Incredible worlds, credible results. *Erkenntnis*, 70:119–131.

Kuorikoski, J., Lehtinen, A., and Marchionni, C. (2010). Economic modelling as robustness analysis. *British Journal for Philosophy of Science*, 61:541–567.

Lange, M. (2014). On "Minimal Model Explanations": A reply to Batterman and Rice. *Philosophy of Science*, 82:292–305.

Laudan, L. (1971). William Whewell on the consillience of inductions. *The Monist*, 55(3):368–391.

Lavis, D. A. (2008). Boltzmann, Gibbs, and the concept of equilibrium. *Philosophy of Science*, 75(5):682–696.

Lehtinen, A. (2013). On the impossibility of amalgamating evidence. *Journal for General Philosophy of Science*, 44(101):101–110.

Lehtinen, A. (2016). Allocating confirmation with derivational robustness. *Philosophical Studies*, pages 1–25.

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in the History and Philosophy of Biology and Biomedical Sciences*, 44(4):503–514.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54(4):421–431.

Levins, R. (1993). A response to Orzack and Sober: Formal analysis and the fluidity of science. *The Quarterly Review of Biology*, 68(4):547–555.

Li, L.-F. (2012). Introduction to renormalization in field theory. In Henley, E. and Ellis, S., editors, *100 years of subatomic particle physics*, pages 1–31. World Scientific, London.

Lilliard, A. S., Lerner, M. D., Hopkins, E. J., Dore, R. A., Smith, E. D., and Palmquist, C. M. (2013). The impact of pretend play on children's development: A review of the evidence. *Psychological Bulletin*, 139(1):1–34.

Lisciandra, C. (2016). Robustness analysis and tractability in modeling. *European Journal for Philosophy of Science*, pages 3–17.

Lotka, A. J. (1925). *Elements of Physical Biology*. Williams and Wilikins, Baltimore.

Macy, M. W. and Willer, R. (2002). From factors to actors: Computational sociology and agent-based modellng. *Annual Review of Sociology*, 28:143–166.

March for Science (2017). About us. https://www.marchforscience.com/-mission/.

Marcoci, A. and Nguyen, J. (2017). Scientific rationality by degrees. In Massimi, M., Romeijn, J.-W., and Schurz, G., editors, *EPSA 2015 Selected Papers. European Studies of Philosophy of Science*, pages 85–97. Springer, Frankfurt.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., Matschoss, P. R., Plattner, G.-K., Yohe, G. W., and Zwiers, F. (2010). Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. *Intergovernmental Panel on Climate Change*, pages 1–7.

Mastrandrea, M. D., Mach, K. J., Plattner, G.-K., Edenhofer, O., Stocker, T. F., Field, C. B., Ebi, K. L., and Matschoss., P. R. (2011). The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. *Climate Change*, 108:675–691.

Matthews, D. (2010). The housing crisis: What can we learn from history? *BBC History Magazine*, pages 1–5.

Mattingly, D. (2005). Modern tests of Lorentz invariance. *Living Reviews in Relativity*, 8:1–84.

McCloskey, D. N. (1993). Other things equal: The a-prime/c-prime theorem. *Eastern Economic Journal*, 19(2):235–238.

McMullin, E. (1985). Galilean idealizations. *Studies in the History and Philosophy of Science*, 16(3):247–263.

Montuschi, E. (2009). Questions of evidence in evidence-based policy. *Axiomathes*, 19:425–439.

Morgan, M. G. (2014). Use and (abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America*, 111:7176–7184.

Morgan, M. S. (2013). *The World in the Model: How Economists Work and Think*. Cambridge University Press, Cambridge.

Morreau, M. (2014). Mr. Fit, Mr. Simplicity and Mr. Scope: From social choice to theory choice. *Erkenntnis*, 79:1253–68.

Morreau, M. (2015). Theory choice and social choice: Kuhn vindicated. *Mind*, 124(493):239–262.

Morrison, M. and Morgan, M. S. (1999). Models as mediating instruments. In Morrison, M. and Morgan, M. S., editors, *Models as Mediators*, pages 10–37. Cambridge University Press, Cambridge.

Muldoon, R. (2007). Robust simulations. *Philosophy of Science*, 74:873–883.

Muldoon, R., Smith, T., and Weisberg, M. (2012). Segregation that no one seeks. *Philosophy of Science*, 79(1):38–62.

Musgrave, A. (1981). Unreal assumptions in economic theory: The f-twist untwisted. *Kyklos*, 34:377–387.

Mäki, U. (2013). On a paradox of truth, or how not to obscure the issue of whether explanatory models can be true. *Journal of Economic Methodology*, 20(3):268–279.

NICE (2006). The guidelines manual. *Website National Institute for Health and Clinical Excellence (http://www.nice.org.uk)*.

Odenbaugh, J. and Alexandrova, A. (2011). Buyer beware: Robustness analysis in economics and biology. *Biology and Philosophy*, 26:757–771.

Okasha, S. (2011). Theory choice and social choice: Kuhn vs. Arrow. *Mind*, 129(477):83–115.

Okasha, S. (2015). On Arrow's theorem and scientific rationality: Reply to Morreau and Stegenga. *Mind*, pages 1–16.

Orzack, S. H. and Sober, E. (1993). A critical assessment of Levins's the strategy of model building in population biology. *The Quarterly Review of Biology*, 68(4):533–546.

Parker, W. (2011). When climate models agree: The significance of robust model predictions. *Philosophy of Science*, 78(4):579–600.

Pierce, C. S. (1868). Some consequences of four incapacities. *Journal of Speculative Philosophy*, 2:140–157.

Pincock, C. (forthcoming). Accommodating Explanatory Pluralism. In Reutlinger, A. and Saatsi, J., editors, *Explanation Beyond Causation*. Oxford University Press, Oxford.

Pivato, M. (2014). Formal utilitarianism and range voting. *Mathematical Social Sciences*, 67:50–56.

Putnam, H. (1975). *Mathematics, Matter and Method*. Cambridge University Press., Cambdrige.

Raerinne, J. (2013). Robustness and sensitivity in biological models. *Philosophical Studies*, 166:285–303.

Reif, F. (1985). *Statistical and Thermal Physics*. McGraw-Hill, Michigan.

Reiss, J. (2008). *Error in Economics: Towards a More Evidence-Based Methodology*. Routledge, London.

Reiss, J. (2012). The explanation paradox. *Journal for Economic Methodology*, 19(1):43–62.

Reiss, J. (2013). The explanation paradox redux. *Journal for Economic Methodology*, 20(3):280–292.

Reiss, J. (2015). A pragmatist theory of evidence. *Philosophy of Science*, 82(3):341–362.

Resnik, M. D. (2011). *Choices: An introduction to decision theory*. University of Minnesota Press, Minneapolis and London.

Rice, C. C. (2016). Factive scientific representation without accurate representation. *Biology and Philosophy*, 31:81–102.

Rickles, D. (2007). Econophysics for philosophers. *Studies in History and Philosophy of Modern Physics*, 38:948–978.

Rizza, D. (2014). Arrow's theorem and theory choice. *Synthese*, 191:1847–1856.

Roberts, B. (2013). When we do (and do not) have a classical arrow of time. *Philosophy of Science*, 80(5):1112–1124.

Rol, M. (2013). Reply to Julian Reiss. *Journal of Economic Methodology*, 20(3):244–249.

Romer, P. M. (1993). Two strategies of economic development: Using ideas and producing ideas. *Proceedings of the World Bank annual conference on development economics*, pages 63–91.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Sackett, D., Rosenberg, W., Gray, M., Haynes, B., and Richardson, S. (1996). Evidence-based medicine: What it is and what it isn't. *British Medical Journal*, 312:71–72.

Schelling, T. C. (1971). Dynamic models of seggregation. *Journal of Mathematical Sociology*, 1:143–1861.

Schickore, J. and Coko, K. (2013). Using multiple means of determination. *International Studies in Philosophy of Science*, 27(3):295–313.

Schupbach, J. (2016). Robustness analysis as explanatory reasoning. *British Journal for Philosophy of Science*, pages 1–26.

Sen, A. (1977). On weights and measures: Informational constraints in social welfare analysis. *Econometrica*, 45:1539–1572.

Sen, A. (1986). Social choice theory. In Arrow, K. J. and Intriligator, M. D., editors, *Handbook of Mathematical Economics*, pages 1073–1181. Elsevier, Amsterdam.

Singer, D. and Singer, J. L. (1990). *The house of make believe: Children's play and the developing imagination*. Harvard University Press, Cambridge.

Smith, N. (2016). Data geeks are taking over economics. *Bloomberg View*, pages 1–4.

Smith, P. K. and Dutton, S. (1979). Play and training in direct and innovative problem solving. *Child Development*, 50(3):830–836.

Smith, R. and Rennie, D. (2014). Evidence based medicine: An oral history. *British Medical Journal*, pages 1–8.

Sober, E. (1989). Independent evidence about a common cause. *Philosophy of Science*, 56(2):275–287.

Socolow, R. H. (2011). High-consequence outcomes and internal disagreements: Tell us more, please. *Climate Change*, 108:775–790.

Stefanovich, E. V. (2014). Relativistic quantum dynamics. https://arxiv.-org/abs/physics/0504062.

Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, 76:650–661.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(4):497–507.

Stegenga, J. (2012). Rerum concordia discors: Robustness and discordant multimodal evidence. In Soler, L., Trizio, E., Nickles, T., and Wimsatt, W. C., editors, *Characterizing the Robustness of Science*, pages 207–226. Springer, Frankfurt.

Stegenga, J. (2013). An impossibility theorem for amalgmating evidence. *Synthese*, 190:2391–2411.

Stegenga, J. (2014). Herding QATs: Quality assessment tools for evidence in medicine. In Hunemann, P., Lambert, G., and Silberstein, M., editors, *Classification, Disease and Evidence: New Essays in the Philosophy of Medicine*, pages 193–211. Springer, London.

Stegenga, J. (2015). Theory Choice and Social Choice: Okasha versus Sen. *Mind*, 124(493):263–277.

Strevens, M. (2004). The causal and unification accounts of explanation unified causally. *Nous*, 38(1):154–176.

Strevens, M. (2008). *Depth*. Harvard University Press, Cambridge MA.

Strevens, M. (2017). Notes on Bayesian Confirmation Theory. *Manuscript*, pages 1–151.

Sugden, R. (2000). Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology*, 7(1):1–31.

The Economist (2016). The post-truth world: Yes, I'd lie to you. *The Economist*, pages 1–11.

Thompson, E., Helgeson, C., and Frigg, R. (2016). Expert judgment for climate change adaption. *Philosophy of Science*, 83(5):1110–1121.

Thébault, K., Bradley, S., and Reutlinger, A. (2016). Modelling inequality. *British Journal for Philosophy of Science*, pages 1–23.

Troffaes, M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45:17–29.

UN (2015). Conference of the parties twenty-first session Paris 30 November to 11 December 2015: Adoption of the Paris Agreement. http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf.

University of the West of England (2008). The history of council housing. https://fet.uwe.ac.uk/conweb/house-ages/council-housing/print.htm.

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–60.

Weber, M. (2011). Experimentation versus theory choice: A social-epistemological approach. In Schmidt, H. B., Sirtes, D., and Weber, M., editors, *Collective Epistemology*, pages 1–26. Ontos Verlag, Frankfurt.

Weed, D. L. (2005). Weight of evidence: A review of concept and methods. *Risk Analysis*, 25(6):1545–1557.

Weisberg, M. (2006a). Forty years of 'the strategy': Levins on model building and idealizations. *Biology and Philosophy*, 21:623–645.

Weisberg, M. (2006b). Robustness analysis. *Philosophy of Science*, 73(5):730–742.

Weisberg, M. (2007). Three kinds of idealizations. *The Journal of Philosophy*, 104(12):639–659.

Weisberg, M. (2013). *Simulation and Similarity: Using models to Understand the World*. Oxford University Press, Oxford.

Weisberg, M. and Reisman, K. (2008). The robust Volterra principle. *Philosophy of Science*, 75(1):106–131.

Werndl, C. and Frigg, R. (2015). Reconceptualizing equilibrium in Boltzmannian statistical mechanics and characterizising its existence. *Studies in History and Philosophy of Modern Physics*, 49:19–31.

Werndl, C. and Frigg, R. (2016). When does a Boltzmannian equilibrium exist? In Timpson, C., editor, *Quantum Foundations of Statistical Mechanics*. Oxford University Press, Oxford.

Wimsatt, W. (1987). False models as means to truer theories. In Nitecki, M. and Hoffman, A., editors, *Neutral Models in Biology*, pages 23–55. Oxford University Press, Oxford.

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In Brewer, M. B. and Collins, B. E., editors, *Re-Engineering Philosophy for Limited Beings*, pages 124–159. Jossey-Bass, San Francisco.

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

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, 13(2):219–240.

Wootton, D. (2007). *Bad Medicine: Doctors Doing Harm Since Hippocrates*. Oxford University Press, Oxford.

Ylikoski, P. and Aydinonat, E. (2014). Understanding with theoretical models. *Journal for Economic Methodology*, 21(1):19–36.