Jaakko Kuorikoski

Society by Numbers:
Studies on Model-Based Explanations in the Social Sciences
Table of Contents

Acknowledgements

List of Original Publications

Part I: Introductory Essay
1. The aim and scope of this study ......................................... 11
2. The fall of theory and the rise of the model ....................... 12
3. Doing things with models .................................................... 15
4. Explanation: from covering laws to causation and constitution ................................................................. 21
5. Social sciences: the necessity of naturalism ..................... 26
6. The need for theories of the middle range ....................... 28
7. On philosophical method and the division of cognitive labour ................................................................. 31
8. Overview of the articles ........................................................ 33
9. Concluding remarks .............................................................. 42

Part II: Explaining with Equilibria and Mechanisms

Dissecting Explanatory Power .............................................. 53
Petri Ylikoski and Jaakko Kuorikoski

Explaining With Equilibria ...................................................... 81
Jaakko Kuorikoski

Two Concepts of Mechanism:
Componential Causal System and Abstract Form of Interaction ................................................................. 103
Jaakko Kuorikoski

Part III: Explanation on the Boundaries of Economics

Unrealistic Assumptions in Rational Choice Theory ....... 131
Aki Lehtinen and Jaakko Kuorikoski

Economic Modelling as Robustness Analysis ...................... 161
Jaakko Kuorikoski, Aki Lehtinen and Caterina Marchionni

Computing the Perfect Model: Why Do Economists Shun Simulation? ................................................................. 195
Aki Lehtinen and Jaakko Kuorikoski
Acknowledgements

In my experience, philosophy certainly has been nothing like the solitary pursuit of contemplative truth that many people take proper philosophizing to be. This research has not taken place in hermetic isolation from other people and the rest of the visible, audible and tangible world. I am grateful for all the energetic and social contexts in which philosophy contained within these pages has been collectively produced. Throughout my career as a philosopher, I have had the privilege of being a part of an active research group as well as of a broader philosophical community. The writing of this dissertation has never felt like a momentous struggle, something that I would be glad to be done with. This is no doubt due to the fact that I have always been able to work within a vibrant community, simply doing research rather than trying to produce monolithic personal existential statement of vast proportions.

I am most indebted to my supervisor Petri Ylikoski. Petri introduced me to the emerging constellation of people that has now evolved into the Philosophy of Science Group and I consider myself Petri’s student in more ways than one. Besides sharing with him many views about specialized topics within the philosophy of science, I have also learned many priceless lessons about what philosophy in general ought to be and how it should be pursued. Petri is also the co-author of what is possibly the theoretically central piece in this collection of articles. Petri has tirelessly offered me advice on academic practicalities and tricks of the trade related to conferencing, application writing and teaching.

By far the biggest influence on the content of this dissertation has been Aki Lehtinen, since he has been my co-author in three out six of the following articles. Aki has been invaluable in keeping my feet on the ground an insisting that my sometimes far-fetched and usually sketchy ideas have to be presented in a way understandable to other people and preferably backed up by actual arguments. It goes without saying that without Aki’s in-depth knowledge of economics and political science, this dissertation would have been something far less than it is now. Aki has also been a delightful travelling companion in the numerous conference trips during which we have presented the fruits of our joint labour.

The third person to have contributed to this dissertation in a very concrete fashion is Caterina Marchionni. In addition to
the knowledge and insight directly contributed to our paper on robustness analysis, Cate’s keen observations have also greatly contributed to most of the other papers. I am also thankful for the fact that Cate has willingly chosen to endure the wintertime darkness and the often iffy cuisine in order to enrich and enliven our research group.

I am extremely grateful for Matti Sintonen for being a constant source of support and an inspiration for the philosophical study of explanation for me as well as for the Finnish philosophy-community in general. Matti has been instrumental in creating an environment in which philosophy of science can be effectively pursued. Without Matti’s infectious optimism, I might never have set foot on this particular path. Matti has also guided me successfully through the difficult final stages of this project.

A very special thanks goes out to my boss Uskali Mäki, for his support, inspiration and constructive criticism. Uskali has never let an imprecise formulation slip through unnoticed and his constant demand for more thorough analysis has been a vital ingredient in the making of this dissertation. Most importantly, Uskali is largely responsible for the fact that I have been able to work within such a wonderful research group. Uskali’s vision and ambition are invaluable driving forces without which most of the achievements of our philosophy of science community would not have taken place.

I would also like to thank Marja-Liisa Kakkuri-Knuuttila for all the enjoyable years working and teaching in the former Helsinki School of Economics, now part of the Aalto University. Marja-Liisa was kind enough to hire me as a research assistant when I was still an undergraduate and working with her has taught me immensely about doing the kind of philosophy that is relevant outside our narrow specialty and especially about communicating constructively with people from various academic backgrounds. I am still not convinced that Aristotle got everything right, however. I am greatly indebted to all my colleagues within the Philosophy of Science Group, Emrah Aydınonat, Tarja Knuuttila, Inkeri Koskinen, Alessandro Lanteri, Pekka Mäkelä, Jani Raerinne, Anna-Mari Rusanen, Päivi Seppälä and Till-Grüne Yanoff, but special thanks go to Tomi Kokkonen and Samuli Pöyhönen. The endless (in many ways) discussions with Tomi, starting from the very first days of being new students of philosophy, have probably influenced my
thinking more than anything else during all these years spent in
the university. Having Samuli as an office-mate for the last couple
of years has greatly enriched my daily working existence and I
am looking forward to actually starting all the joint projects now
planned during our spontaneous brain-storming moments.

My philosophical existence has for a good while been cross-
departmental (and cross-institutional) and I owe gratitude to a
number of people in both departments. Panu Raatikainen has
always been willing to help me out in every conceivable situation
(including practicalities related to this book) and I greatly value
his philosophical knowledge and friendship – even though I
proudly consider myself an arrogant philosopher of science.
Ilkka Niiniluoto’s support was crucial in the early phases of my
post-graduate studies and he has been a constant inspiration as a
paradigm of a great philosopher. Other philosophical comrades
that I have had the privilege of knowing and without whom this
dissertation would undoubtedly look quite different include Jussi
Backman, Raul Hakli, Antti Kauppinen, Vili Lähteenmäki, Ville
Paukkonen, Tuukka Tanninen, and Pilvi and Teemu Toppinen. I
thank you all. I would also like to thank the administrative staff
of both philosophy departments, Ilpo Halonen, Auli Kaipainen, Terhi
Kiiskinen and Karolina Kokko-Uusitalo, for all their help. From the
philosophy unit at Aalto University, I thank Kaisa Heinlahti, Jukka
Mäkinen and Kristina Rolin. There are probably others who would
deserve mentioning here but now escape my memory.

Particular thanks are due to the preliminary examiners,
Johannes Persson and Erik Weber, whose comments and kind
words were of great help in finishing this dissertation. The Finnish
Cultural Foundation has funded a major part of the work that has
gone into this study and I am privileged in having been employed
in three research projects funded by the Academy of Finland while
working on this thesis. Thanks also to Tinde Päivärinta for doing the
layout of this book.

But by far the biggest debt of gratitude I owe to my parents
Erkki and Ulla. They have never questioned my unorthodox choice
of occupation, have always supported me in all ways imaginable,
and are largely responsible for who I am today. The final thank-you
goes to Maria for her love.

Helsinki, March 2010
Jaakko Kuorikoski
List of original publications

This dissertation consists of the following publications:

I Dissecting Explanatory Power
Petri Ylikoski and Jaakko Kuorikoski

II Explaining With Equilibria
Jaakko Kuorikoski

III Two Concepts of Mechanism: Componential Causal System and Abstract Form of Interaction
Jaakko Kuorikoski

IV Unrealistic Assumptions in Rational Choice Theory
Aki Lehtinen and Jaakko Kuorikoski

V Economic Modelling as Robustness Analysis
Jaakko Kuorikoski, Aki Lehtinen and Caterina Marchionni
*British Journal for the Philosophy of Science* (forthcoming)

VI Computing the Perfect Model: Why Do Economists Shun Simulation?
Aki Lehtinen and Jaakko Kuorikoski
Part I: Introductory Essay

1. The aim and scope of this study

The aim of this dissertation is to provide conceptual tools for the social scientist for clarifying, evaluating and comparing explanations of social phenomena based on formal mathematical models. The dissertation is composed of six related but, in principle, independent research articles. The focus is on relatively simple theoretical models, not statistical models. Thus, this dissertation is not primarily about solving specifically philosophical problems, such as the nature of explanation or causation in general. A specific stance towards these questions is presupposed rather than directly argued for. It is assumed that the contrastive counterfactual theory of explanation and the invariance-under-interventions account of causation are correct, but few explicit arguments are given in favour of these theories. The primary condition of success for these studies is whether social scientists themselves can find something of use in them - something that would, in the end, enable them to provide better explanations of social phenomena. Whether the theses put forth in these studies adhere to the intuitions of philosophers is, at most, a secondary issue.

The most important philosophical framework used in these studies is the contrastive-counterfactual theory of explanation, as developed by James Woodward (2003). This is especially true for the first part of the dissertation, but the contrastive counterfactual theory of explanation is also used or presupposed in many of the arguments in the second part as well. Although few explicit arguments are given for this philosophical framework (for these, see ibid; Ylikoski 2001), it is hoped that these studies themselves can be seen as an argument in favour of it. The best argument for a philosophical theory concerning scientific practice is that it helps to improve that practice. If and when the contrastive-counterfactual theory of explanation helps us to elucidate actual explanations and explicate our implicit standards of evaluating them, then the theory has proven its mettle. But before the reader can assess whether this is indeed the case, some remarks on the place and history of models and explanations within the tradition of the philosophy of science as well as social sciences are in order.
The structure of this introduction is as follows. First, the current centrality of the concept (or, more likely, the word) ‘model’ in philosophy of science is placed within its historical discursive context. I suggest that the current abundance of model-talk is the result of two partially independent developments: the development of the semantic view of theories and the rise of the philosophies of the special sciences. In section 3, the questions of representation and the epistemic value of models are discussed and my own stances on these issues are also presented. Although most of the theses concerning model-based explanation in the articles are relatively independent on these issues, I feel that the reader is better off being fully aware of my philosophical commitments with respect to the epistemology and ontology of modelling. Then some remarks on the history of the theory of explanation are made and the contrastive-counterfactual theory of James Woodward is briefly presented in section 4. In sections 5 and 6, the stances taken on the nature of modelling and explanation are brought to bear on the fundamental issues in the methodology of the social sciences. After this, I briefly reflect on the philosophical methodology used in this dissertation. Finally, I go over the main results of the articles and draw some conclusions.

2. The fall of theory and the rise of the model

The studies in this dissertation are about model-based explanations in the social sciences. The predicate model-based is here meant to be contrasted with the idea that explanations could be directly derived from a theory. This is not purely a terminological matter or a reflection of the current indisputable model-boom in the philosophy of science. The studies comprising this dissertation are about explanations that are not simply a matter of deriving something particular from the more general. The explanations discussed within these pages essentially involve manipulating some external inferential apparatus - a model.

Problems relating to the use of models currently constitute one of the most central areas in philosophy of science. This has not always been the case. There was a time when the default bearer of scientific content was taken to be the theory, and questions of explanation, confirmation and realism were discussed in terms of
the relationships between theory, observations and the world. The positivist conception of the structure of scientific knowledge at the same time gave theory a central role as the locus of scientific knowledge, and denied that the content of that theory could be anything more than the economic re-organization of occurrent or even directly observable regularities. Theory was conceived as a (preferably axiomatizable) set of sentences couched in a theoretical vocabulary which was connected to observational language through a set of bridge principles. This was the syntactic or ‘received’ view of theories: The content of scientific knowledge is to be conceived as a set of propositionally structured representations anchored to empiria or to the world itself by analytically true correspondence rules or, after this view proved to be untenable, by empirical bridge laws (see e.g. Hempel 1965; Nagel 1961; Suppe 1977).

The philosophical developments of the 1950s and 1960s created difficulties for the syntactic view, mostly because the view was about sentential content and was thus tied to matters of language in general. The idea of analytically true bridge principles was effectively dismantled by Quine’s critique of analyticity (1951), Hanson (1958) pointed out the inescapable theory-ladenness of observations, and it was recognized more generally that the relationship between theoretical content and observations was much more complicated than could be handled with the logical tools of the received view. To overcome these challenges, Patrick Suppes (1960) proposed that theories should be presented by directly specifying the corresponding classes of models, which are mathematical, not sentential, entities. From this suggestion grew the semantic conception, which, more or less, began to identify theories with classes of models, which in turn are understood as (uninterpreted) mathematical structures or trajectories in some state space (see e.g. van Fraassen 1980; Giere 1988; Suppe 1989). Moreover, as Suppes and, later, Ronald Giere have emphasized, even models are usually not tested against naked observations, but models of data (Suppes 1962; Giere (forthcoming)). Thus it can be argued that the core general philosophy of science discourse concerning the form of scientific content has been turning to a concept of model (in a logico-mathematical sense) due to its internal dynamics.

Another factor in the current popularity of model-talk in the philosophy of science literature has been the decreasing overall
importance of the use of formal logical machinery and what might be called problems of general philosophy of science. Instead, what have taken the central stage are the philosophies of the special sciences, which address problems closer to the praxis of science. And that praxis is usually highly model-centred. As philosophy of science has become more closely aligned with the practice of science, models are no longer seen only as logical tools in the rational reconstruction of the content of science, but as important objects of theorizing. Most of our knowledge of evolutionary biology, ecology, psychology, economics and even physics can not be moulded without great violence to fit the format of an axiomatizable set of universal laws. Instead, most actual, workable scientific knowledge, discounting the distorting simplifications of schoolbook or popular science, is packed into more local and context-dependent representations, built opportunistically from components available in one’s toolbox, from empirical hunches and purely ad hoc mathematical tricks as much as from generally accepted theoretical principles. These representations are also called models.

In scientific practice, models are often seen as mediating general theoretical principles and empirical applications (Morrison and Morgan 1999). Thus, if one is interested in the relationship between scientific knowledge and the world, one should look at the practice of constructing and using models and, as has been noted in the literature, the relationship between the construction and use of these inferential apparatuses is always, to some extent, autonomous from general theoretical principles (ibid.). In fact, much of the current practice-oriented discussion on models criticizes the semantic view of linking models too closely to theories (a critique that ignores the differences in the aims of the proponents of the semantic view and the philosophers who study the praxis of model-building). Instead, models always incorporate, and are thus constrained by, factors that cannot be derived from or even argued for on the basis of either theory or data. Models contain context-dependent operationalizations, metaphors, specific empirical assumptions and mathematical techniques introduced purely for tractability. These ingredients mean that model-based explanations have to be investigated by looking at how the models are used in practice, not whether a proposition of a certain kind can be derived from a set of theoretical postulates.
3. Doing things with models

Analogical and thought-experiment-like reasoning is probably as old as reasoning itself. Lucretius ruminated on the nature of the magnet on the basis of possible hypothetical mechanisms based on the atomic theory of matter. Gassendi conceived a molecular model of inheritance of ontogenic information between parents and offspring. Galileo reasoned about the nature of movement on the basis of thought-experimentation on a nonexistent frictionless plane. Yet the use of models as an explicit and distinct research strategy really gained momentum only in the nineteenth century, largely as a result of the need to postulate and reason about unobserved theoretical entities in a tractable but rigorous manner (Hesse 2000).

The word ‘model’ originally referred primarily to concrete objects, like mechanical models, composed of, for example, movable bars, cords, wheels and rollers, or to moulds made of wax as well as physiological (anatomical) models made of plastic. Indeed, in his famous encyclopedia entry on models, Ludwig Boltzmann precludes maps, charts, musical notes or figures from categorization as models, since models ‘always involve a concrete spatial analogy in three dimensions’ (1911). Boltzmann did not require, however, that models should actually exist as physical objects, for they could also be conceived as mental constructs.

The semantic view is a theory about the logical structure of all scientific content, but not all scientific practices involve explicit modelling. Modelling is the activity of constructing and manipulating a thing in order to study some other thing. What characterizes theoretical modelling is a certain epistemic dynamic making use of surrogate reasoning: one first builds something or sets something up, then investigates the properties of that constructed thing, and then ponders how the discovered properties of the constructed thing relate to the real world. Recently, Michael Weisberg (2007) and Peter Godfrey-Smith (2006) have usefully articulated this widely held view of specifically model-based reasoning as the strategy of indirect representation. They claim that instead of directly abstracting some salient aspects of data or a target system into a workable and more systematic scientific representation (strategy of

---

1 This section draws on Kuorikoski and Lehtinen (2009) as well as Knuuttila and Kuorikoski (forthcoming).
direct representation), modellers seek to understand the real world through the procedure of constructing and analyzing hypothetical systems, in other word models. A modeller begins to attack a problem by coming up with a set of simple theoretical principles which, when combined, might be expected to solve the problem (such as providing an explanation for a puzzling phenomenon).

This surrogate system view captures an essential feature of specifically model-based reasoning, but taking it too literally risks creating (and in the philosophy of science literature has indeed created) an additional problem: if model-based reasoning involves two steps, first reasoning about the properties of the model and then from these properties to properties of things in the world, just what properties of the model justify the latter inferential step? The syntactic view took truth or truthlikeness – a semantic relationship between propositionally structured representations and the world – to be the epistemically important property of scientific content. When models were identified as non-sententially structured systems of representation, philosophical thought turned to other relationships between the model and the world that could possibly ‘explain’ the epistemic properties of models and provide grounds for the second inferential step from models to the world. A variety of different morphisms (isomorphism [van Fraassen 1980], homomorphism [Bartels 2006], partial isomorphism [French 2003]) have been proposed to fulfil this role, but none have exhibited the right formal properties or saved all the central intuitions so as to be considered as ‘The’ relationship between models and the world. Other proposed concepts, such as similarity (Giere 1988; 2004), seem too vague to have much analytic or explanatory potential (Suárez 2003). Similarity is simply a placeholder, not an actual explanatory factor. The same diagnosis can be made of Robert Sugden’s characterization (2002) of ‘credibility’ as the grounds for making the inductive leap from the world of models to the world itself.

Some pragmatic approaches to representation (dis)solve some of the problems of the semantic notion of representation mentioned above by the introduction of the users’ intentions, which create the directionality needed to establish a representative relationship (Giere 2004; Giere forthcoming; Mäki 2009). But this also comes at a price. When representation is grounded primarily on the specific goals and representing activities of humans as opposed to the facts
about the representative vehicle and the target object, nothing very substantive can be said about the relationship of representation in general. Another danger in stressing the intentions of the modellers is the conception of representation as a mental act – as if intentionality inside the head could magically create a relationship between a model and the world. This way of conceptualizing aboutness – conceiving intending as a special mental act that a-causally endows utterances with their meanings or things with their representational content – was identified as a form of mythical thinking and subsequently refuted by Wittgenstein (1953). It is also worth avoiding in the context of models.

Mauricio Suárez (2003; 2004) has gone farthest in arguing for a minimalist account of representation which resists saying anything substantive about the supposed basis on which the representational power of representative vehicles rests (that is, whether it rests, for instance, on isomorphism or similarity). Instead, Suárez builds his inferential account of representation directly on the idea of surrogate reasoning: i.e. the model represents something in virtue of its capacity to lead a ‘competent and informed user to a consideration of the target’, and the right kind of constitution to allow agents to correctly draw inferences from it correctly (Suárez 2004). While the aforementioned representational capacity of a model is created and maintained by the inferential activity of representation-users, the stipulation concerning the right kind of constitution saves the intuition that what can be a model of a certain system is not completely arbitrary. What kind of inferences a cognitive agent with certain sensory apparatus and cognitive capacities can make with the help of an external model-thing depends on the causal properties of the model-thing (see also Vorms 2009). Suárez’s account can thus be seen as an application of Robert Brandom’s view of representation in general: it is the inferential properties of objects (in relation to the agents using them) that constitute their representational properties and appealing to any primitive representational concepts to explain inferential properties would, therefore, put the cart before the horse (Brandom 1994).

Conceiving models as independent entities and construing their representational capacity on the basis of the kind of inferences that can be made with them paves the way for considering modeling simply as a form of extended cognition, as extended inference in
which the relevant cognitive unit of analysis is the user-model pair (or the research-community-model pair). In a paper not included in this dissertation, Aki Lehtinen and I (2009) denied that any philosophical account of representation (or of other model-world mappings, for that matter) can explain the epistemic value of models. This is because there is nothing to be explained: modelling is simply inference-making with the help of external cognitive aids (such as a formal language or a diagram). Inferring here means, roughly, using formal syntactic rules to derive contentful expressions from other contentful expressions in a truth- or probability-preserving manner. This means that the only epistemic questions concern the reliability of the assumptions and the reliability of the inferences. What sets modelling apart from pure thought experimentation is that in the former the inferences from assumptions to conclusions are not conducted entirely in the head of the modeller or only in natural language, but rather with the help of external inferential aids such as diagrams, mathematical formulas or computer programmes. As de Donato Rodriguez and Zamora-Bonilla (2009) put it, models function as inferential prostheses. What is doing the cognitive work in modelling is not the individual, but the individual-model pair. Modelling is essentially inference from assumptions to conclusions conducted by an extended cognitive system (Giere 2002; Kuorikoski and Lehtinen 2009).

However, as the idea of modelling as indirect representation makes clear, modelling certainly is distinct from ordinary inference and argumentation in that we seem to find out genuinely new things by manipulating or investigating an artificial construct. The experience of discovering novel information is common to both modelling and experimentation. This analogy between the epistemic dynamics of modelling and experimentation (cf. Mäki 2005) can be misleading, however. The sense of novelty in modelling is only the result of the essential use of external inferential aids. Using mathematics, diagrammatic reasoning carried out with pen and paper, or computer simulation involves manipulations of representations external to the mind of the human subject, and he or she may not experience this manipulation as inference-making, i.e. as something phenomenologically similar to thinking. What the human subject experiences is more akin to experimentation with an artefactual, abstract or imaginary system. The modeller manipulates
graphs or mathematical equations – something external to his or her mind – and then finds something new about the abstract object that is represented by the equations or graphs. Yet, from the perspective of the whole extended cognitive system consisting of the modeller and the external representations (the model), there is no experimentation, only inference. The only ‘epistemic access’ (cf. Mäki 2009) that the extended cognitive system has to the target system is via the original causal connection that was required for the formulation of the substantial empirical assumptions from which the inferences are made. The things that are found out are new only in the sense that the conclusions were not transparent to the unaided reasoning powers of the modeller.

A possible objection to identifying models with concrete external inferential apparatuses is that it amounts to conflating models with individual model-descriptions. Most often scientists take the identity of a given model to be tied to an abstract, partially interpreted, mathematical object (e.g. a harmonic oscillator), which can be described in multiple different ways. However, although modelling necessarily involves abstracting, models in themselves are not abstract entities. Abstraction is an activity performed by a cognitive agent, but the end result of that activity, the abstraction of something, need not in itself be an abstract entity. Instead, it is (often) a material thing used to represent something. It is usually the inferential rather than the material properties of these abstractions that are epistemically important for the modeller. Although the material means of a representation often do matter in subtle ways for what inferences can be made with it (Vorms 2009), the aim in modelling is to minimize or control for these influences: if a conclusion derived from a model is found to be a consequence of a particular feature of a material representation lacking an intended interpretation, the conclusion is deemed to be an artefact without much epistemic value. Therefore, it often makes perfect sense to further abstract from multiple individual representations to their common inferential properties and then label these common inferential properties as ‘the’ model itself. These inferential properties are, of course, not intrinsic to the representations, but rather depend on the context in which they are used.

There is, thus, no need to abandon the distinction between ‘the model’ and its various descriptions (cf. Mäki 2009). For example,
many kinds of public representations facilitate similar kinds of
inferences from spring constants and amplitudes to total energy,
and this makes all of these representations models of ‘the’ harmonic
oscillator. Such abstractions are often extremely useful in co-
ordinating cognitive labour. By referring to them, we refer only to
a set of inferences and can therefore disregard the material things
that enable us to make these inferences in practice. The material
form these representations may take is usually not relevant to the
epistemic problems at hand: whether a differential equation was
solved on a piece of paper, blackboard, or computer is not usually
relevant to whether or not it was solved correctly. This is why it
is natural to think that the ‘identity’ of the model of the harmonic
oscillator resides precisely in these common inferential properties
of the various material representations, i.e. in the abstract object.
Nevertheless, we should resist reifying these abstractions as abstract
objects in themselves.

The conception of modelling as extended inference also
helps us to dismiss one common misconception about modelling,
namely the idea that modelling automatically entails a preference
for quantitative data or quantifiable issues over a more qualitative
approach. The function of the mathematics is not to relate numbers,
but to facilitate secure derivations from theoretical principles and
thus to enable reasoning about matters that are too complex for
natural language (Simon 1957, Ch. 6). Most social systems consist
of multiple interdependencies and are frequently characterized
by mutual consistency conditions involving multiple factors, such
as equilibrium constraints. This makes them next to impossible to
theorize about in natural language, which by its very nature can
only reliably facilitate sequential inferences. Thus, despite the title
of this dissertation, much of theoretical modelling in the social
sciences is about the qualitative features of the social systems
modelled and need not involve any explicit statistics. The principal
job of theoretical models is to keep our reasoning straight, not to
mirror numerically exact laws of society.

What kind of model is appropriate depends on the goals of
modelling; some types of models are obviously better suited to some
specific purposes than others. Social scientific models can be used to
suggest explanations for certain specific or general phenomena; to
carry out virtual experiments; to specify and even help to execute
policies based on a model; to design new institutions; to make predictions; to conduct thought experiments using a model; to derive solutions to theoretical problems; to explore the limits and range of possible outcomes consistent with questions that can be answered using a model; to develop concepts and classificatory systems; or simply as a pedagogical aid. Simple and analytically tractable models are usually taken to be good at conceptual elaboration, theory development and explanation. Among others, Boyd and Richerson (1987) argue that replacing unintelligible phenomena with unintelligible models does not increase our understanding and that simple models are thus appropriate for explanation. However, their uses can be limited. For example, Grün-Yanoff (2009) takes a rather extreme position in suggesting that simple or ‘minimal’ models can only be used to disprove pre-theoretic impossibility intuitions (such as that racial segregation cannot arise from non-racist preferences). Simple models (such as the IS-LM model in macroeconomics) are also usually well suited for pedagogical purposes, although complex simulations can also be used to provide illustrative examples. Relatively simple atheoretical associational models are often superior to more complex theoretical models in pure prediction tasks. On the other hand, policy analysis is often conducted with the aid of extremely complicated and data-rich computer simulations that are used to predict the consequences of possible interventions (for example country- or city-specific epidemiological models and computational macroeconomic models used by central banks). The focus in the following essays is mostly on relatively simple theoretical models that aim to explain some puzzling societal phenomenon. Hence, the next issue to be explored is the contested nature of explanation.

4. Explanation: from covering laws to causation and constitution

One important function of theoretical models is to provide explanations. But should the social sciences be in the business of providing explanation in the first place? One of the central methodological disputes within the social sciences has been about the alleged difference between interpretation and explanation and whether the social sciences should even strive to be explanatory.
The argument has been something like the following: since human action is not susceptible to universal laws, but its central constituents (beliefs and desires of the subjects) are immediately accessible to the social scientist through the process of empathic identification, interpretation, not explanation, is the road to greater understanding of society. One of the key ingredients in this debate has thus been the idea that explanation, as opposed to interpretation, is about subsuming singular events under universal laws. The question to ask now is whether the current views on the nature of explanation support this presupposition.

Measured by the number of publications, theory of explanation has been one of the most productive sub-fields of the philosophy of science for the last half a century. A part of the explanation for this striking fertility has to be the existence of a particularly good target at which all the discussants can aim their criticism: the deductive-nomological model. The D-N model states that explanation amounts to a derivation of the explanandum-event from a set of initial conditions and at least one universal, exceptionless law (Hempel 1965). The D-N model allowed the theory of explanation to flourish as subsequent counterexamples gave rise to new refined analyses which themselves prompted new counterexamples. Although the myriad of counterexamples has resulted in the almost universal abandonment of the D-N model (see, e.g., Woodward 2003, 154-161), the form and methodology of this debate has left the theory of explanation largely stipulative in the sense that it has ignored the question of why we want explanations in the first place; explanation just is the laying out of the causal history of an event (Salmon 1984) or unification of our overall worldview (Kitcher 1989), or whatever seems to save the most pre-theoretical intuitions and historical explanatory improvements.

The current contender to the throne, the contrastive counterfactual account of James Woodward (2003) used in the studies comprising this dissertation, is a marked improvement on this situation. According to Woodward, explanation consists of tracing or exhibiting functional dependency relations between variables. Explanation is thus doubly contrastive; the functional relationship links the possible values of the explanans to possible values of the explanandum. These explanatory relationships provide understanding by giving answers to what-if-things-had-been-
different questions (w-questions) concerning the consequences of counterfactual or hypothetical changes in the values of the explanans variable. These answers are the basis of inferential performance constitutive of understanding. In the case of causal explanations or explanations given on the basis of a causal model, the relevant hypothetical changes to consider are interventions, ideally surgical manipulations that affect only the explanans variable of interest and leave the rest of the model intact (apart from the changes caused by the change in the explanans variable dictated by the model, of course). Causal explanations thus trace dependencies that are invariant under interventions. Woodward’s theory is essential for the first part of the dissertation, but a broadly contrastive-counterfactual view of explanation is also presupposed in much of the argumentation in the second part.

Woodward’s account intimately links explanatory knowledge to our capacity to function in the world as goal-directed manipulators. Whereas the epistemic conception of explanation behind the D-N-model was exclusively concerned about expectability and predictive power, thus denying that explanatory information had any deeper import, Woodward’s theory stresses our role as active agents, as opposed to merely passive observers of events. The correlate concept of understanding refers to the ability to make correct inferences on the basis of received knowledge and causal knowledge concerns the effects of manipulations and therefore licenses inferences about the effects of our actions. Understanding thus lies not only in the correctness of inference, but sometimes also in the effectiveness of action based on the information to be understood. As is stated in the first article, understanding should not be conceived as a special method or a mental state. Understanding is not a state-concept to begin with, but a concept akin to an ability attributed according to manifest performances (Ylikoski 2009). The criteria of understanding are public and the role of the concept is regulatory: attribution of understanding signals reliability of inference and action with respect to the object of understanding.

An important point concerning philosophical methodology is that the analytical import of the contrastive-counterfactual theory is not dependent on anybody’s intuitions about what is and what is not explanatory. The theory makes a crucial distinction between the kinds of inferences that can be legitimately made from purely
descriptive information and from causal/explanatory information: purely descriptive information does not license inferences beyond what has actually happened or what will likely happen if the system is left unperturbed. Whether one calls providing information with the appropriate modal implications explanation or exclamation, and whether the attributions of explanatory power derived from the contrastive-counterfactual theory match all pre-theoretical intuitions, does not really matter. Hence, even in the unlikely event that some details of Woodward’s account were to be proven wrong in the future, much of the analytic work done with the contrastive-counterfactual theory of explanation in this dissertation should still remain valid.

Woodward’s theory is a one of causal explanations. It is a realist theory in the sense that the exhibited invariant causal dependencies are a feature of the world and the existence of these dependencies is independent of our ways of conceptualization and representational practices. Explanation is also factive in that an explanation is genuine only when it is true, i.e., only when the required causal dependency obtains and the cited values of the variables are (sufficiently) correct. These points are made here in order to alleviate the possible concern that stressing the inferential performance as constituent of understanding might render the approach somehow ‘instrumentalist’. Explanations refer to mind-independent things in the world and we can collectively be mistaken about whether an explanation is correct. Beyond this basic realist commitment, the contrastive-counterfactual theory is not really committed to any specific metaphysics of causation. The invariance-under-interventions account of causation is committed to the broadly accepted idea that causal dependencies are always (or at least outside fundamental physics) realized by a mechanism (and mechanistic ideas are certainly a key ingredient in this dissertation as well) but there are alternative metaphysical ways of cashing out this mechanistic idea (Glennan 2005; Machamer et.al. 2000). The choice of fundamental metaphysics for causation should not matter that much for analyses of explanations made on the basis of the interventionist account.

However, explaining something is still an epistemic activity, conducted by limited cognitive agents such as ourselves. Hence, explanations always relate things in the world conceptualized in
some way or another and our cognitive limitations set constraints on what and how we can reason about the *relata* of explanations. Thus, the goodness of an explanation depends also on how easily a limited cognitive agent can use the explanatory information to make relevant counterfactual inferences. Explicating the contrast class for an explanandum is a way of making the conceptualization of the explanandum explicit, and thus an object of argument and debate. The fact that explaining something is an epistemic activity also means that the explanatory information conveyed in an explanation has to be accessible to the receiver of the information – otherwise any improvement in inferential performance would be a complete mystery. Providing an actual explanation and claiming that there is one, somewhere to be found, are therefore not the same thing. As Woodward has argued, this simple enough point makes untenable theories of explanation that render the explanatoriness of a single explanation relative to either the whole belief system (unification), or to a yet to be found ideal explanatory text (e.g. Railton 1981) (Woodward 2003, 179-181). Of course, this does not mean that a single scientist ought to be in the possession of all relevant information, all the time, in order to be judged to understand something. This would, in most cases, be beyond the cognitive powers of humans and require the idea that a single explanatory speech act or text passage could magically hold much more information than it actually does. The constraint of epistemic accessibility should be read as demanding that the explanatory information should be retrievable and usable by a relevant scientific sub-community, such as a research-group, in reasonable time. One should always keep in mind that scientific cognition is massively distributed and extended and that much less happens within individual heads than is usually thought in traditional philosophy of science.

Causal dependency is not the only ontic dependency-relation supporting explanations. Many mechanistic explanations relate the causal properties of the parts of the mechanism (together with their organization) to a property of the whole. This relation of dependency is not causal, since it is not a process in time relating ontologically separate entities. Instead, it is a relation of constitution. The atomic structure of an elastic solid does not cause the object to be elastic since the atomic structure is not a distinct entity, or a set of distinct properties, from the elasticity of the object. Thus
far, constitutive explanation has received relatively little attention, Cummins 1983 and Craver 2007 being notable exceptions. Luckily, the contrastive-counterfactual theory can also be generalized to cover constitutive explanation. Constitutive explanations answer w-questions concerning what would happen to a property of the whole if the properties or organization of the parts were changed (intervened on) in some way. The relevant conceptions of invariance and modularity have to be tweaked somewhat, since the constituted property cannot be intervened on independently of its constituting base (one cannot causally change the elasticity of an object without at the same time changing the properties of its atomic structure), but this is not a pressing concern here. Many model-based explanations in the social sciences are actually constitutive explanations, or at least will be argued as such in the following articles.

5. Social sciences: The necessity of naturalism

The use of models is often associated with and legitimized by the nomothetic ideal of science that stresses the use of general or universal laws and predictive power as hallmarks of a truly scientific enterprise. Understandably, this frequent appeal to the nomothetic ideal may also be used as a reason to argue for the futility of modelling in areas where there seem to be no universal covering laws to be had, such as the social sciences. The nomothetic ideal of science is presupposed in much of the classical as well as current methodological discussion concerning idiographic/interpretive and nomothetic/functionalist paradigms of social research. It is clear that if explanation amounted to the subsumption of events under universal laws, then there would have been very few explanatory successes in the history of the social sciences and scant hope for any in the future as well.

Woodward’s theory of explanation is in many ways antithetical to the nomothetic ideal of science, and the conception of understanding as inferential ability accommodates both causal explanation as well as the products of interpretation and ascriptions of meaning. Universality and exceptionlessness are not the modal properties that make a relationship or generalization explanatory. Explanations trace dependencies, not occurrence regularities, and
generality is an epistemic and pragmatic virtue, but not itself constitutive of explanatory power. There might be a place for a substantial conception of a law of nature in philosophy of science, but that place is in the interpretation of fundamental physics, not the sciences of hierarchical complexity such as the biological sciences or, arguably, the social sciences. However, what the conception of understanding does require is that social scientific explanations pick out relations of either causal or constitutive dependence and these dependencies are features of the systems being investigated, not features of our conceptualizations or interpretations. Cataloguing the interpretations and meanings that the subjects attach to their social context is purely descriptive unless these meanings can be embedded in relations of causal or constitutive dependence. The same goes for intentional states attributed to individual agents: folk psychological rationalizations of behaviour are explanatory only if they pick out the true causes of behaviour, something that, it had been different, would have resulted in different behaviour as well. Thus, although the concept of explanation used here does not make use of the concept of natural law, it does support weak methodological naturalism in that, insofar as the social sciences should strive for explanations in the first place, they are not characterized by any special method of verstehen incompatible with other forms of scientific explanation.

The contrastive-counterfactual theory of explanation and especially the importance of the concept of intervention also provide a strong argument for theoretical (in contrast to purely atheoretical empiricism) and explanatory (in contrast to purely interpretive) social science: only causal-explanatory knowledge allows us to change the society in a goal-directed manner. This brings us back to the emergence of the social sciences as distinct academic disciplines. The positive sciences of the society were founded upon the idea that empirical laws governing suicide, crime, poverty or mental illness could be discovered and that these laws could be used to maintain and improve the health of the newly created national states. Although the fruits of science-based societal engineering have proven to be much more modest than originally hoped for,
the fundamental point still stands: without explanatory ambitions, social sciences lose their policy relevance and their broader societal relevance becomes suspect.

The conception of understanding as inferential performance is deflationary in that it denies that understanding has a deeper essence from which the public manifestations of understanding flow. This means that vague intuitions concerning intelligibility and appeals to the psychological sense of understanding can, in principle, always be calibrated or altogether replaced by making the implicit criteria of understanding explicit (Ylikoski 2009). Many methodological controversies and quarrels between disciplines and social scientific research traditions trade on unarticulated notions of what is a better explanation. Much of this bickering could be avoided and fruitful and constructive dialogue made possible if these vague intuitions about explanatory goodness and explanatory relevance were replaced by explicating what kinds of things are meant to be explained and how this explaining is to be done. The articulation of what can and cannot be explained with a given model, based on analyzing what kind of counterfactual inferences the model justifies, can be seen as the central contribution made in the articles comprising this dissertation.

6. The need for theories of the middle range

As mentioned above, it can be argued that the sciences have become more and more model-centred. Mary Hesse (2000) argues that this trend started already in the 19th century. To some extent, this is also true of the social sciences. The first steps of the social sciences as distinct scientific disciplines were inspired by, and to some extent founded on, the newly invented statistical apparatus of the 19th century. The 19th century also gave us the Edgeworth-box, von Thünen’s isolated state (see, e.g., Mäki forthcoming) and the germs of the idea that economic knowledge should be moulded into a similar format as that of physical knowledge (Mirowski 1989). However, even economics did not really become thoroughly model-based until around the middle of the 20th century, and the only truly widespread use of models in the other social sciences is still statistical modelling, purely descriptive as well as causal. Yet many
things are called models in the social sciences. Often an analytically useful conceptualization of phenomena represented with a 2-2 field or a hypothesis of simple causal connection represented by a boxes and arrows diagram is called a model. In these essays, the focus is limited to models in a stricter sense, to what might be called theoretical formal models. Theoretical modelling follows the strategy of indirect representation. By theoretical models, I mean external inferential devices that are used to infer non-trivial consequences from prior, usually theoretical, assumptions, not to infer estimates of theoretical constructs from masses of data. By non-triviality, I mean that the inference essentially involves some kind of manipulation of the external inferential apparatus, usually syntactic manipulation of symbols according to well-defined syntactic rules. Thus, most theoretical models are indeed ‘formal’ or mathematical, but graphical reasoning and purely physical devices, such as the famous hydraulic economy of A.W. Phillips, qualify also. The focus on theoretical models also means that I am not addressing statistical models of either a descriptive or causal kind. Although statistical models always involve prior theoretical assumptions, they are still essentially data-driven in the sense that the conclusions of interest are about estimates of some parameter or other, not about what follows from a set of prior assumptions. Statistical modelling instantiates the strategy of abstract direct representation, not indirect representation (cf. Weisberg 2007). Especially statistical causal modelling involves completely different philosophical issues from those of theoretical modelling and would thus deserve a separate dissertation.

In philosophy of science, models have been conceived as mediating between theory and observation or between theory and the world. In the social sciences, modelling can be seen as a potential answer to another mediation problem, namely what Robert Merton famously described as the need for theories of the middle range. Merton accused ‘grand’ sociological theorizing as too abstract and general to of any use in understanding specific empirical phenomena. Grand theories á la Marx or Talcott Parsons serve only to conceptualize empirical social phenomena and do not actually exclude any relevant possibilities. Thus, grand theories do not offer any grounds for making inferences as to what might happen or might have happened and are for this reason non-explanatory. On the other hand, banal empiricism, either in the guise of blindly running...
regressions (variable sociology) or cataloguing the qualitative interpretations or meanings that the subjects attach to their social context, is purely descriptive and thus also non-explanatory (Merton 1957, section II). As has been claimed by the proponents of the recent analytical sociology movement (e.g., Hedström 2005), formal modelling can be seen as a promising methodology for producing such middle-range accounts. Although no single study in this dissertation includes an explicit argument for more widespread use of modelling in the social sciences, the dissertation as a whole can be seen as a supporting statement for modelling methodology in the social sciences.

Although economic modelling is the most prominent form of formal modelling in the social sciences, it is by no means the whole story. The boundaries between the different social sciences have traditionally been drawn according to the scale and nature of the forces they study. Sociology studies the non-market driven societal or macro phenomena in industrial and post-industrial societies, economics (arguably) phenomena governed by markets, social psychology interaction in relatively small groups etc. These traditional ways of defining the division of labour in the social sciences do correspond to some extent to differences in their respective modelling practices. Economists’ theoretical practices nowadays consist nearly entirely of modelling with only a relatively limited set of modelling methodologies, whereas few sociologists rely on modelling, but those who do use a more heterogenous set of model templates (Edling 2002).

However, modelling frameworks cross disciplinary boundaries and an argument could be made that the traditional way of dividing the disciplines should be, and is actually in the process of being, replaced by a division of labour according to the use of different modelling tools (Humphreys 2004, 71). Most models in social sciences are indeed more or less of the off-the-shelf type, i.e. abstract structures that can be applied with varying degrees of adjustment to systems with intuitively very different causal make-ups. The principles of stochastic processes, network models or constraint optimization are similar regardless of who or what (individuals, groups, countries, ideas…) is doing the random walking, networking or maximizing. However, as will become apparent later, this does not mean that there are no important discipline related differences in the ways in
which these templates are applied. For example, sociologists and economists conceptualize, use and interpret simulation models in importantly different ways and these differences reflect the deep methodological and substantial differences in their respective approaches to social phenomena.

7. On philosophical method and the division of cognitive labour

Scientific research reports always include a section devoted to methodology. Interestingly enough, this is not true of every, or even most, treatises in philosophy. If there is a distinct methodology used in the studies comprising this dissertation, it is not philosophical conceptual analysis – at least if conceptual analysis is understood as the formulation of explicit accounts of some pre-theoretically understood concept which are then tested against pre-theoretic or ‘expert’ intuitions of philosophers (Kitcher 1992). Although these studies do not use philosophical conceptual analysis, the points being made are largely conceptual in nature. Although, as Quine has taught us (1951), no such unique and universal line can be drawn between conceptual and empirical content that would neatly classify every possible statement across all time as being of one kind or the other, conceptual and empirical content can and should be distinguished locally (as a good example, see Burian et. al. 1996).

One only has to acknowledge that all definitions are in principle always open to revision and that the final criterion of ‘truth’ of any conceptual claim is whether its acknowledgement removes confusion and decreases errors in reasoning within the inferential practices in which the concept is actually used. Thus any conceptual claims should not be primarily tested against the intuitions of the philosopher, but against arguments that are in the end empirically motivated and, ultimately, against whether or not the conceptual claims are of any help in the practice of science itself.

This does not mean that scientists could never err in conceptual matters or in their self-assessment of what they are doing. Contrary to the widespread misconception, even thorough naturalism does not preclude a normative, evaluative stance on the scientific practices studied. In my view, philosophical accounts should first and foremost guard against faulty reasoning and a philosophical
account can judge a specific modelling practice, explanation or an expression of momentary self-reflection made by a scientist to be in error as long as the grounds for the judgement include something more than the mere intuitions of the philosopher. For example, the drawing of a conceptual distinction may help in identifying fallacious inferences that arise from using the same word in two different inferential contexts. And whether these inferences are fallacious or not is grounded, in the end, not on platonic meanings or essences, but on the inevitable errors that arise from mismatches between our linguistic and non-linguistic practices and the causal structure of the world. The fact that philosophy and the sciences are on the same continuum does not mean that the former has no say on the validity of particular claims made in the latter sphere – quite the contrary. Philosophy can take part in the scientific debate precisely because it is, or to the extent that it is, part of science. What should count in science as well as in philosophy are the arguments, and arguments are evaluated on the basis of the credibility and relevance of the reasons and the validity of the argument form, not on the basis of the department the maker of the argument happens to be affiliated with.

Most of the data used in these studies is not from the unfortunately under-researched area of explanatory cognition (although especially the first essay is informed by it), nor from systematic surveys of what scientists themselves take their models to explain, nor from ethnographic field studies investigating how scientists actually build their models. Although thoroughly naturalistic, these essays are not about the psychology or self-understanding of social scientists. If these studies were primarily about scientific cognition, then empirically exploring ‘cognition in the wild’ would be the way to proceed (cf. Hutchins 1995). In contrast, the theses argued in these studies are about the content of science, and the primary data are therefore scientific research articles and the explicit arguments used by the scientists themselves. The content of science is the public corpus of knowledge and methods distributed within the scientific community and on the pages of

---

3 The trouble with much of philosophy is that the reasons, which are usually either semantic (‘we would not call something x, if it did not…’) or modal (‘it would be impossible for x to be y if it did not…’), are often neither credible nor relevant.
scientific journals, not the mental states or psychological processes within the heads of individual scientists.

Another feature these studies share with their more scientific cousins is that most of them are collaborations. This is not a defect but a virtue. Since the content of philosophy does not qualitatively differ from that of the special sciences, neither should its method of production. The romantic lay-conception of philosophers (and probably a self-conception of a good many philosophers as well) has been of a lone heroic mind, solitarily exploring the fundamental nature of reality or human existence. But science is a fundamentally social enterprise and scientific cognition is massively extended and distributed. What is truly constitutive of scientific progress is not what happens inside any individual head, but the collaborative effort in laboratories and on the pages of scientific journals. The same should be true of philosophy as well. The philosopher does not have any privileged access to The Truth and there is no psychological evidence suggesting that the belief system of an individual (even of a philosopher) is usually more coherent than that of a collective (such as a research group of a field). Just as any other well-developed field of knowledge production, philosophy should embrace the gains of division of labour, which, after all, is the very foundation of modernity.

8. Overview of the articles

The remainder of this dissertation is divided into two parts. The first part develops the general conceptual tools with which to understand and evaluate model-based explanations in the social sciences. Although the primary intended application is the social sciences, the first three articles are essentially general philosophy of science. The first article lays down the theory of explanation used throughout and develops the different dimensions of explanatory power. The next two articles explore two concepts central to social science modelling: equilibrium and mechanism. The final part of this dissertation applies these concepts and ideas to specific modelling techniques in the social sciences. The fourth article discusses the impact of unrealistic assumptions on the validity of explanations based on rational choice models, and the fifth article explores the use
of robustness analysis in gauging and ameliorating this impact in economics. The sixth article argues that the resistance to theoretical agent-based simulations in economics is partly due to the specific, though implicit, conception of understanding held by economists.

I Dissecting Explanatory Power

The claim that some model or theory provides a better explanation than another is often used as an argument in favour of that model or theory and disputes between approaches and whole disciplines are often couched in terms of explanatory power. But these disputes and arguments invariably fail to articulate the used criterion or criteria of the goodness of explanation. What does it actually mean to explain something better or worse? Is there a standard or metric according to which we can compare explanations and is this metric different and independent from the comparative credibility or probability of the rival explanations? The first article aims to provide answers to these questions and thus lays down the conceptual tools used to discuss various models and explanations in the rest of the articles.

‘Dissecting Explanatory Power’ uses the contrastive-counterfactual theory to articulate five dimensions of explanatory power: precision, non-sensitivity, factual accuracy, degree of integration and cognitive salience. The guiding principle behind this taxonomy is simple: The more correct inferences to counterfactual situations can actually be made on the basis of the explanatory information (by a cognitively limited human agent), the better the explanation. That the relevant inferences are to concern counterfactual situations means that explanatory information characteristically has implications beyond what actually happened or what will happen if nothing is changed. The requirement that the inferences have to be actually achievable by cognitively limited beings, such as us, is a consequence of treating explaining as a concrete epistemic activity involving real people, not as an end product of completed science of philosophers’ fantasies. These dimensions can be in conflict with each other and there are some systematic trade-offs between them. Thus, different disciplines can give different weights to these aspects of explanatory goodness and thereby systematically value certain kinds of explanations. These explanatory virtues should also
be kept distinct from evidential virtues; whether an explanation is
good is a different question from whether a putative explanation is
likely to be true, and all the explanatory virtues are characterized
with the presupposition that the explanation is (roughly) true.

II Explaining with Equilibria

Most theoretical models in the social sciences are tightly linked to
two concepts: equilibrium and mechanism. The concept of equilibrium
is almost definitional for economic model building, but the majority
of other social scientific models also incorporate some equilibrium
concept or other. This is not a coincidence, a purely empirical claim
or (only) a matter of prevalent conservative socio-political views
modellers have about the nature of society. Equilibrium methodology
is an almost necessary consequence of (at least) two constraints on
formal modelling: the requirement of analytic tractability and the
relative paucity of solid and yet generalizable empirical constraints
on modelling assumptions.

Many modellers in the social sciences, especially economists and
so-called economics imperialists, argue that equilibrium modelling
holds some intrinsic scientific virtues (see, e.g., Lazear 2000). Are
equilibrium explanations then somehow especially virtuous? This is
precisely the kind of question a philosophical theory of explanation
should be good for. Yet despite their importance, equilibrium
explanations have thus far received little philosophical attention. A
notable exception is Elliott Sober’s article ‘Equilibrium Explanation’
(1983), in which Sober argues that equilibrium explanations
constitute a counterexample to the causal theory of explanation,
since the initial condition of an equilibrium system, the causal
history of the equilibrium state, are explanatorily irrelevant for the
equilibrium state. The argument in ‘Explaining with Equilibria’ is
that Sober has misidentified the principal *explanans* of equilibrium
models. Since explanation requires dependency, not regularity
or stability, it is the structural properties (encoded in parameter
values and functional forms), that explain specific properties of
equilibria, rather than the initial conditions explaining the obtaining
of equilibrium, as Sober thinks. This explanatory relationship is
constitutive, not causal, since it relates properties of the parts of the
system to a systemic property of the whole and is not in itself a process in time (although equilibration itself is a process in time, of course). This constitutive dependency can in turn be used to causally explain changes in the equilibrium state by structural changes in the underlying equilibrium mechanism. In terms of modelling, this is the standard practice of comparative statics familiar from textbook economics. The points above suggest that any perceived epistemic advantages of equilibrium explanations arise from the fact that modelling the underlying equilibrating mechanism is often possible with only relatively modest assumptions about the constituent parts of the mechanism, not from the fact that we can be ignorant about the actual initial conditions of the system. This means that the equilibrium dynamics usually do play an explanatory role, since the explanatory structural assumptions about the causal properties of the mechanism components are often embedded in the arguments for the adopted dynamics.

III Two Concepts of Mechanism: Componential Causal System and Abstract Form of Interaction

Theoretical models and middle range theories in the social sciences are usually conceived as being about social mechanisms. Social mechanisms are thought to be relatively stable constellations of social actions that generate statistically discernible regularities. Social mechanisms are also precisely what Merton’s middle range theories should capture. As argued in ‘Explaining with Equilibria’, equilibrium models, insofar as they are about causal relations to begin with, are essentially models of mechanisms and the primary explanatory relationship is that of constitution between the properties of the parts and a property of the whole. When pressed about the concept of social mechanism, analytically minded social theorists usually refer to something akin to methodological individualism (see the essays in Hedström and Swedberg 1998). The concept of mechanism seems to refer to little more than to the idea of explaining something macro with something more micro, to the idea that this explanation has something to do with the causal properties of the micro and to some account of sequentiality or process.

The concept of mechanism is also enjoying something of a boom in the philosophy of science literature. The ‘new mechanists’ have
offered mechanistic accounts of explanation, research strategies, inter-field relations and metaphysics of causation (Bechtel 2006; Bechtel and Richardson 1993; Craver 2007; Glennan 1996; Machamer et. al. 2000). The concept of mechanism seems to fit the sciences of multi-layered hierarchical complexity (such as cellular biology) better than the concept of law borrowed from the philosopher’s conception of fundamental physics. However, the mechanisms discussed in this philosophical literature bear little resemblance to the social mechanisms described by analytic sociologists, for example. The question thus remains whether this philosophical literature has any bearing for methodological issues concerning mechanistic reasoning in the social sciences. In ‘Two Concepts of Mechanism’, I distinguish between two conceptions of what it is to be a mechanism, which are based on two very different ways of mechanistic theorizing and empirical research. The new mechanists take mechanisms to be componential causal systems, in which the intrinsic causal properties of ontologically distinguishable parts realize the functional or system-level properties of the mechanism according to their organization. The idea is that a quintessential mechanism is a contrivance serving a function, itself composed of a number of parts serving different causal roles. Exploring such mechanisms proceeds along the heuristics of decomposition and localization discussed in the literature on mechanistic research in the life sciences (Bechtel and Richardson 1993).

In the social sciences, the concept of mechanism should not be taken so literally. Social mechanisms are better seen as structurally similar forms of agent interaction, what I call abstract forms of interaction. In contrast to componential causal systems, our understanding of such mechanisms is not necessarily improved by leaning more about the intrinsic causal properties of the most obvious parts of the social systems (individuals), since the causal properties relevant for the system level property are relational, rather than monadic. Thus, the way to gain understanding is to reason about the consequences of different forms of interaction by building simple and idealized models, and only combining these models to reason about the properties of constellations of multiple mechanisms (cf. Boyd and Richerson 1987). However, both concepts of mechanism have similar explanatory virtues familiar from the literature on mechanistic explanation. It should also be stressed
that the distinction does not match neatly any imagined boundary between the natural and human sciences.

Since both concepts of mechanism are associated with distinct research heuristics, their use carries the possibility of introducing characteristic biases to the research and explanations utilizing them. The typical biases associated with the componential causal system -conception include many reductionistic fallacies discussed by William Wimsatt (2007), such as misguidedly attributing a relational property to be an intrinsic property of a component part or presupposing too much context-insensitivity to the causal properties of the component parts. Abstract form of interaction -concepts (AFI) may induce other kinds of biases, such as automatically attributing further causal properties to the component parts of an AFI-mechanism on the basis that the component parts of some other similar AFI-mechanism also exhibit them. An exclusive focus on only certain kinds of AFI-concepts, perhaps motivated by some ideal of cross-disciplinary unification, can also lead to myopic research, which misses important causal properties of the systems under investigation. This danger is clear in the imperialistic applications of economic ideas that use market metaphors to explain all kinds of societal phenomena.

IV Unrealistic Assumptions in Rational Choice Theory

Theoretical models in the social sciences are almost invariably highly unrealistic. The multi-dimensional complexity of societal phenomena makes idealizations and other falsehoods necessary if the models are to be tractable or possess any degree of generality. The realism of assumptions has been a much discussed methodological issue as long as there has been modelling. Most of this discussion in the social sciences has concentrated on the set of idealizations that are almost definitiona of economic models, i.e., the assumptions depicting the rational economic man.

In ‘Unrealistic Assumptions in Rational Choice Theory’, Aki Lehtinen and I argue that much of the methodological discussion concerning rational choice theory is misguided in that it presupposes that there is a unitary and substantial rational choice theory underlying all its models. We argue instead that rational choice
models can explain in different ways and that the empirical and explanatory content is usually not found in the rationality axioms, but in the operationalization arguments used to transfer the abstract machinery of constrained optimization and equilibrium into empirically interpretable models. When looking at the realism of the assumptions, one should thus first look at the way in which the self-interest assumption is actually interpreted in the model as behavioural or motivational assumptions. The reason why rational choice seems to work better in intuitively economic matters is that these operationalization arguments are usually much more credible in straightforwardly economic contexts than in non-economic ‘imperialistic’ applications.

Second, whether making unrealistic assumptions invalidates the explanations made on the basis of the corresponding model depends on whether the model-result is dependent on something that is obviously wrong in these assumptions. Explanation is about tracing dependencies and if the explanatory dependency (the model-result) itself is not dependent on some particular falsehood in the assumptions, that particular falsehood does not matter for the model’s explanatory value. Thus when Woodward’s theory of explanation is applied to rational choice models, three distinct kinds of explanantia can be discerned. First, some rational choice models can indeed be seen as formalized and usually aggregated intentional (folk-psychological) explanations in that the self-interest assumption is interpreted as telling what the motivations of the agents are supposed to be and that the model result (explanation) is dependent on the correctness of this assumption. Second, some rational choice models should be interpreted as (legitimately) claiming only that the agents act as if they were maximizing their self-interested utility. If some other factors (like selection) make it plausible that the aggregate behavioural consequences are consistent with this assumption, the accuracy of the psychological attribution does not matter. Third, sometimes even the individual behavioural assumptions are not what drive the result. Some well-known rational choice modelling results demonstrate such highly abstract and generic systemic properties, that they are largely robust with respect to what the agents are supposed to know, want or behave. These may not be the most interesting of properties of social systems, but accusing such models on the basis that they make unrealistic
assumptions concerning the agents is not the right way of criticizing
the modelling practice in question.

V Economic Modelling as Robustness Analysis

As Uskali Mäki has suggested, some idealizations are introduced
into models in order to isolate a specific causal tendency from
the clutter of real world phenomena (Mäki 1992; 1994). However,
many falsities in models are there because of the requirements of
mathematical tractability – the specific assumption could not be
included in the mathematical model in a more realistic manner
in a way that would have left the model to still be analytically
solvable (Hindriks 2006). Models are representations built using
formal languages and this means that assumptions have to often
be implemented in the model in a more specific form than would
really be justifiable on empirical grounds. For example, there might
be reason to believe that an interesting economic phenomenon is
dependent on transportation costs, but no reason to think that the
particular form of transportation costs should matter, and perhaps
no data showing what the transportation costs look like in potential
areas of application. However, the costs have to be implemented
somehow and the model might be conveniently solvable using
a particular functional form for the transportation costs – a form
which happens to be known to be false.

‘Economic Modelling as Robustness Analysis’ argues that the
peculiar practice of model refinement in economics, in which the same
well-known results are repeatedly derived using slightly different
alternative modelling assumptions, can be, somewhat charitably,
interpreted as a process of collective derivational robustness analysis
– a strategy of testing which false assumptions are actually fatal for
the common modelling results or theorems. The article uses William
Wimsatt’s account of robustness analysis (2007, chpt. 4) to provide an
epistemic rationale for this practice by showing how the procedure
guards against the inevitable falsities in modelling assumptions.
Robustness is here understood in an epistemic sense, as a property
of our representational and inferential apparatuses rather than as
a causal property of the investigated systems. Robustness analysis
gauges the extent to which we can get away with not knowing all
the details and still understand the mechanism. This means that sticking to roughly the same modelling framework may not be an irrational (or ideological) bias, distorted view of human nature or even a metaphysical prejudice, but a reasonable response to the paucity of data and the complexity and heterogeneity of the studied phenomenon. The article also provides an argument showing how derivational robustness analysis can rationally change our beliefs about the world without bringing in new observations; if our degree of belief in different modelling assumptions varies, then learning that a specific modelling result does not depend on particular false assumptions, rationally raises our confidence about the modelling result.

**VI Computing the Perfect Model: Why Do Economists Shun Simulation?**

Tractability assumptions are usually necessitated by the requirement that the model can be solved analytically, i.e., that the equilibrium conditions, the equilibrium state or some relevant partial derivatives can actually be analytically derived from the model, rather than inferred by computing the model output using a range of different model inputs or parameter values. Analytic solutions have been highly regarded especially in economics, which has historically been averse to computational methods in general and to agent-based micro simulations in particular. This is somewhat peculiar, especially compared to the other social sciences, in which agent-based simulations have been relatively prominent (with respect to the use of any theoretical models) ever since computational methods have been widely available. Computational methods are becoming ever more prevalent in economics, but they are still mostly used to compute equilibrium paths of models structured in the same way as the old analytic equilibrium models (e.g. computational dynamic stochastic general equilibrium models), rather than true agent based simulations. Do analytic solutions have some epistemic properties that make them attractive to the extent that the consequent costs in model realism are justified? ‘Computing the Perfect Model: Why Do Economists Shun Simulation’ provides a tentative answer: not really. The article catalogues the epistemic virtues attributed to analytic solutions, such as generality, epistemic transparency and
Comparability with respect to other models, and differentiates between the possible senses in which the computer can be said to be an epistemically opaque black box. However, an advocate of true bottom-up simulation techniques can, in principle, answer all these criticisms.

As a tentative hypothesis for the lack of enthusiasm of the economists, the article suggests that simulations do not fit well with the particular conception of scientific understanding shared by economists: the ideal of the perfect economic model. The very training and the admission criteria to the top theoretical journals tell of an implicit conception of understanding in which the very (cognitive) act of deriving the model solution is a constitutive part of understanding of the model and, consequently, of economic phenomena. This implicit conception is compared to Philip Kitcher’s theory of scientific explanation (1989; 1993), according to which deriving descriptions of phenomena to be explained with the help of a set of fixed argument patterns also constitutes the cognitive element in the concept of scientific understanding. Kitcher’s idea that reasoning and explaining is about the application of argument schemas fits well with economists’ practice of using the same principles of utility maximization and equilibrium to explain diverse phenomena. However, from the viewpoint of the contrastive-counterfactual theory and the associated conception of understanding as inferential ability, this derivational conception and the associated ideal of unification are on the wrong tracks. Thus, economists’ resistance to agent-based simulation is at least partly misguided (although there are a number of legitimate causes for distrusting simulations independent of the misguided conception of understanding).

9. Concluding remarks

What can we take home from these studies? In practice, evaluations and comparisons of explanations are almost always based on vague and implicit criteria – on unarticulated and idiosyncratic intuitions. These intuitions are the product of accidents of personal intellectual histories and disciplinary indoctrination, not of an explicit and reason-based discussion about the epistemic goals of a particular scientific
enterprise. Yet judgements of explanatory relevance and goodness of explanation are constantly being used in debates concerning the relative merits of rival theories or whole research approaches and paradigms. This situation is especially prominent in the social sciences. On the one hand, explanations based on equilibrium concepts, explanations referring to underlying mechanisms in general or individual action in particular or explanations based on analytically solvable models are simply taken to be better or more ‘scientific’ than some alternatives. On the other hand, explanations based on familiar folk-psychological concepts and couched in a narrative form are almost guaranteed to elicit the psychological sense of understanding and thus give an impression of providing especially deep insights into social phenomena. A situation in which major epistemological principles driving method and theory choice remain implicit and beyond reasoned debate is clearly not optimal with respect to the advancement of the social sciences. The aim of these studies is to improve this situation.

These theses are based on an exceedingly simple idea: intuitions about explanatory relevance and explanatory power should be replaced by explicating what is supposed to depend on what. Explanation traces dependencies, not regularities, and the understanding provided by an explanation can be made explicit by enumerating the inferences to counterfactual situations made possible by the knowledge of the explanatory dependency. After these matters have been settled, evaluation of the evidence for the explanatory dependency and the theoretic and pragmatic relevance of the explanatory information can commence – and this time starting from assumptions that have been explicitly discussed and hopefully agreed upon. This central idea is presented and elaborated in the first article ‘Dissecting Explanatory Power’, in which five dimensions of explanatory virtue are differentiated.

Explanations trace dependencies and understanding is competence in counterfactual inference. From these ideas it follows that equilibrium models can explain causally as well as constitutively. However, their explanatory power does not flow from the fact that we can ignore the exact initial conditions nor are they somehow especially scientific. Equilibrium models are useful to the extent that we can constitutively explain interesting systemic properties with only modest assumptions about the causal properties of the
constituent parts. The central ideas also help us to see beyond the intuitive idea that uncovering mechanisms is explanatory. They help us to see how and why descriptions of mechanisms are explanatory and that there is more than one concept of mechanism at play in the sciences, concepts that correspond to different ways of modelling and doing empirical research.

The correct identification of the *explanandum* and the *explanans* – of what depends on what – is of crucial importance in the assessment of social science models. It is not enough to just accuse a model of making unrealistic assumptions, since all models invariably do so. Such criticism is valid only if the explanatory dependency is itself dependent on the false assumptions. As an example, rational choice models are not all alike in this respect and the explanatory factors can differ from model to model. This is something that is missed by most of the critics of rational choice. However, this point by itself is not an argument for rational choice modelling. The use of rational choice templates is in all probability a severely biased research heuristic and other model types should be used to control for its characteristic errors – but pursuing this line of argument goes beyond the scope of this dissertation (but see Ylikoski and Kuorikoski 2008). More generally, the checking of robustness of the modelling result with respect to problematic modelling assumptions is an important part of theoretical modelling. This observation can be used to partially account for the peculiar practice of deriving old and well-known results from slightly different assumptions, which is found in the pages of economics journals. There is a clear epistemic rationale to this practise: modelling is inference and robustness analysis is a means of gauging and sometimes improving the reliability of our model-based inferential practices.

Unarticulated intuitions concerning the very essence of understanding remain a powerful factor driving the choice of methodology. This is apparent in economists’ attitudes towards agent-based simulations, which do not fit the economists’ conception of perfect economic model. For economists, the very act of deriving a conclusion from a small set of privileged economic concepts is the gold standard of economic understanding. But explanations trace dependencies and understanding is competence in counterfactual inference. The criterion of correct understanding should be whether the counterfactual inferences are correct and reliable, not how the
inferences are made (e.g., according to a particular equilibrium argument pattern) or to what extent the reasoning process is offloaded outside individual heads.

But what if explanation is not essentially about dependencies and what if understanding is something other than inferential competence? Even though I have criticized the evidential use of philosophical intuition, is it not the case that my preferred theory of explanation in the end rests on intuitions about what really is explanatory? This is not the case. The reader is free to insist that something else, such as deriving a description of a phenomenon from initial conditions and natural laws, is what explanation is really about, but the conclusions concerning the kinds of inferences the exhibition of invariant dependencies enables still stand. If the preferred alternative essence of explanation or understanding holds some further epistemic value over and above the inferential conception advocated here, the burden of proof is upon the reader to demonstrate this. If the observations made in these studies help, even just a little bit, to make our criteria for good model building and model-based explanations more transparent, then this dissertation as a whole is an argument in favour of the philosophical theories used therein. My ultimate hope is thus that these studies will show that philosophy of science can be relevant for the practise of science.

Bibliography


French, Steven. (2003), “A model-theoretic account of representation (or, I don’t know much about art…but I know it involves isomorphism)”, Philosophy of Science 70: 1472-1483.


Hanson, Norwood (1958), Patterns of Discovery, Cambridge UK: Cambridge University Press.


Morgan, Mary and Margaret Morrison (1999), Models as Mediators. Perspectives on Natural and Social Science. Cambridge: Cambridge University Press.


Surrogate Systems”, *Erkenntnis* 70: 29-43.


