History, Philosophy & Theory of the Life Sciences

Hsiang-Ke Chao
Szu-Ting Chen
Roberta L. Millstein *Editors*

Mechanism and Causality in Biology and Economics

Springer

Mechanism and Causality in Biology and Economics

History, Philosophy and Theory of the Life Sciences

Volume 3

For further volumes:
http://www.springer.com/series/8916

Hsiang-Ke Chao • Szu-Ting Chen
Roberta L. Millstein
Editors

Mechanism and Causality in Biology and Economics

Springer

Editors
Hsiang-Ke Chao
Department of Economics
National Tsing Hua University
Hsinchu, Taiwan

Szu-Ting Chen
Graduate Institute of Philosophy
National Tsing Hua University
Hsinchu, Taiwan

Roberta L. Millstein
Department of Philosophy
University of California, Davis
Davis, CA, USA

ISSN 2211-1948 ISSN 2211-1956 (electronic)
ISBN 978-94-007-2453-2 ISBN 978-94-007-2454-9 (eBook)
DOI 10.1007/978-94-007-2454-9
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013939059

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgments

This volume could not have been published without the help, guidance, and support of many people and organizations. In 2008, in light of the maturity of the study of philosophical questions in biology and economics among scholars in Taiwan, Hsiang-Ke Chao and Ruey-Lin Chen proposed an international conference on the philosophy of biology and economics. With funding from Taiwan's National Science Council, the East Asian academic community mobilized to participate in the 2008 conference on model and evolution in biology and economics. This conference hosted three prominent keynote speakers: Mary S. Morgan, Alexander Rosenberg, and C. Kenneth Waters. Many scholars and students benefited from the academic connections they established at this event. To sustain the dialogue with the international community of related research fields, Hsiang-Ke invited Szu-Ting Chen to arrange the 2011 conference—titled *Taiwan Conference on the Philosophy of Biology and Economics: Mechanism and Causality*—held on March 24–25 at the National Tsing Hua University, Hsinchu City, Taiwan. The conference received financial aid from Taiwan's National Science Council, the College of Technology Management, the Research Center for the Humanities and Social Sciences, and the Research and Development Office at the National Tsing Hua University, along with the Taiwan Economic Association. The chapters in this volume are the result of our conference dialogues and presentations. During the conference, Roberta Millstein was invited to join the book's editorial board because of her expertise in the philosophy of biology. She provided critical expertise in managing publication of the collected articles in her field.

Each article has been evaluated through a double-blind process by two anonymous reviewers, and their comments have proved to be most helpful in guiding the authors' revision of their articles. We extend our gratitude to those cautious and generous anonymous reviewers. We also thank our diligent and hardworking authors. It is not an exaggeration to say that, without the effort of the reviewers and authors, the quality and shape of this volume would have been very different.

Special thanks go to Marcel Boumans and Kevin Hoover for their advices in preparing the publication of this book. Hsiang-Ke Chao would like to thank his colleagues at the Department of Economics at National Tsing Hua University for

their warm support: Hwei-Lin Chuang, Elaine Chyi, Chao-Hsi Huang, Eric S. Lin, Shu-Shuan Lu, Chor-Yiu Sin, and Shih-Ying Wu. Szu-Ting Chen would like to express his gratitude to the colleagues at the Graduate Institute of Philosophy at National Tsing Hua University, especially Chi-Chun Chiu and Chon-Ip Ng, for their continuing support. Finally, we also thank the editors at Springer: the publishing editor Ties Nijssen and his colleague Christie Lue as well as the series editor Thomas Reydon. Their efforts have lent polish and precision to the design and content of this volume.

August 2012

Hsiang-Ke Chao
Szu-Ting Chen
Roberta L. Millstein

Contents

Contributors

Marcel Boumans Amsterdam School of Economics, University of Amsterdam, Amsterdam, The Netherlands

Erasmus Institute for Philosophy and Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

Hsiang-Ke Chao Department of Economics, National Tsing Hua University, Hsinchu, Taiwan

Ruey-Lin Chen Department of Philosophy, National Chung Cheng University, Chia-Yi, Taiwan

Szu-Ting Chen Graduate Institute of Philosophy, National Tsing Hua University, Hsinchu, Taiwan

Carl F. Craver Department of Philosophy, Washington University in St. Louis, St. Louis, MO, USA

Lindley Darden Department of Philosophy, University of Maryland, College Park, MD, USA

Till Grüne-Yanoff Avdelningen för Filosofi, Royal Institute of Technology (KTH), Stockholm, Sweden

Kevin D. Hoover Departments of Economics and Philosophy, Duke University, Durham, NC, USA

Marie I. Kaiser DFG-Research Group "Causation and Explanation", University of Cologne, Cologne, Germany

Roberta L. Millstein Department of Philosophy, University of California, Davis, Davis, CA, USA

Julian Reiss Department of Philosophy, Durham University, Durham, UK

Daniel Steel Department of Philosophy, Michigan State University, East Lansing, MI, USA

David Teira Departamento de Lógica, Historia y Filosofía de la ciencia, UNED, Madrid, Spain

Rong-Lin Wang Department of Philosophy, National Taiwan University, Taipei, Taiwan

Chapter 1
Towards the Methodological Turn in the Philosophy of Science

Hsiang-Ke Chao, Szu-Ting Chen, and Roberta L. Millstein

Abstract This chapter provides an introduction to the study of the philosophical notions of mechanisms and causality in biology and economics. This chapter sets the stage for this volume in three ways. First, it gives a broad review of the recent changes and current state of the study of mechanisms and causality in the philosophy of science. Second, consistent with a recent trend in the philosophy of science to focus on scientific practices, it in turn implies the importance of studying the scientific methods employed by researchers. Finally, by way of providing an overview of each chapter in the volume, this chapter demonstrates that biology and economics are two fertile fields for the philosophy of science and shows how biological and economic mechanisms and causality can be synthesized.

1 Introduction

In the philosophy of science, interest has recently shifted from scientific concepts to scientific practices. That means what really matters to philosophers of science, and what philosophical discussions should be based on, is what scientists actually do and how they do it rather than philosophers' visage of what science is and how

H.-K. Chao (✉)
Department of Economics, National Tsing Hua University,
101, Section 2, Kuang Fu Road, 30013 Hsinchu, Taiwan
e-mail: hkchao@mx.nthu.edu.tw

S.-T. Chen
Graduate Institute of Philosophy, National Tsing Hua University,
101, Section 2, Kuang Fu Road, 30013 Hsinchu, Taiwan
e-mail: stchen@mx.nthu.edu.tw

R.L. Millstein
Department of Philosophy, University of California, Davis,
One Shields Avenue, Davis, CA 95616, USA
e-mail: rlmillstein@ucdavis.edu

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_1,

scientists should do it. This application of Hume's guillotine is one of the prevailing trends in the late twentieth century and is sometimes considered as a kind of *naturalism*. Philosophical naturalism is received in various ways. Despite the opposition of supernaturalistic or a priori explanations, as the name suggests, the main theme of naturalism is to align philosophy of science with science and to pay special attention to scientific methods. A sophisticated investigation of the naturalization of philosophy of science requires addressing the questions of how philosophy is naturalized to a specific science and in what respects particular sciences and philosophies of those sciences are similar to one another (Giere 1999, 2008). In contrast, a broadly defined naturalism, which is widely shared by philosophers of science (even by those who do not identify themselves as naturalistic philosophers of science), suggests a two-way study: It on the one hand focuses on scientific practices that matter to philosophical investigations and on the other hand examines philosophical concepts in terms of scientists' work and the devices they employ.[1]

More importantly, as Ronald Giere (2008) points out, philosophical naturalism in turn implies a *methodological stance*. What Giere means is to characterize naturalism as a method that seeks a naturalistic explanation (Giere 2008, p. 214). However, it can be easily extended to a more general naturalistic program that stresses scientific methods. William Bechtel (2008, p. 8) well describes this type of position by stating that the naturalistic philosophy of science attempts to understand science by addressing the following questions: What are the objectives of scientific inquiry? What methods are used to obtain the results? How are the methods and results of science evaluated? How do value issues impinge on the conduct of science? Since the answers to Bechtel's questions crucially require examining scientific methods, the philosophical perspective offered by naturalism necessarily turns to methodology. This edited volume contributes to such a *methodological turn* in the philosophy of science.

In this edited volume, we specifically investigate mechanism and causality in biology and economics. Why do we target mechanism and causality? Despite the fact that they both stand long as important conceptions in the philosophy of science, causality and mechanism are two main guiding ideas that underlie scientists' practices of making explanations. To identify the characteristics of a scientific explanation, we need first to explore what causality and mechanism are and how scientists infer their existence, then conjoin the discussion of causality with that of mechanism for a comparative study.

We particularly focus on the context of biology and economics for three reasons. First, recent developments in the philosophy of science have shown that the philosophy of biology and economics are two of the most fertile fields. The findings in these subdisciplines not only posit serious challenges to but also provide novel ideas for traditional accounts in the philosophy of science that are based mainly on the physical sciences. Second, the current trend of investigating biological or

[1] This point is also suggested in Bechtel (2008, pp. 8–9). For philosophical investigations of scientific devices, examples are experimental and observational instruments by Ian Hacking (1983), models by Mary Morgan and Margret Morrison (1999), and by the semantic or model-based view philosophers such as Ronald Giere (1988, 1999) and Bas van Fraassen (1980, 1989).

economic issues by employing the concepts and tools developed in the other field (e.g., evolutionary game theory, behavioral economics) has drawn substantial attention among scientists and philosophers of science alike. A study that juxtaposes biology with economics and explores a deeper understanding of various philosophical and methodological issues would prove meaningful. Daniel Steel's (2007) highly acclaimed book has demonstrated this. Finally, recent accounts of mechanism and causality in the philosophy of science are often associated with biology and economics. Whereas the philosophy of mechanism has been developed mainly by philosophers of biology (e.g., Machamer et al. 2000; Glennan 1996, 2002; Bechtel and Abrahamsen 2005), philosophical discussions of causality have been inspired by the practices of economists (e.g., Cartwright 1999, 2007; Woodward 2003). Recent works on causality in economics (e.g., Hoover 2001) have also made significant contributions to current and future research on the methodology of causal structure in science in general. However, even though mechanism and causality occupy the main stage of research in both the philosophy of biology and that of economics, only few studies have been done that bring the accounts in one discipline to the other. This edited volume can be seen as a result of collaborative interaction and mutual understanding among philosophers from different disciplines.

2 Mechanism and Causality in the Philosophy of Science

Although causal inquiry has long been regarded as one of the core elements of science, the focus of the philosophical investigation of causality has changed over time since at least the modern era. Traditionally, the discussion tended to pay much more attention to inquiring about the *metaphysical* aspect of causality. This tendency reached its climax in Hume's famous inquiry about the secret connection between any two events—cause and effect. Then in the first half of the twentieth century, influenced by the positivist philosophy of science and the Humean regularity view of the laws of nature, the discussion shifted to a concern about the *epistemological* aspects of the subject. In particular, attempts have been made to delineate the characteristics of causality by using conditional analysis, that is, by analyzing causality in terms of necessary or sufficient conditions, or both. By temporarily leaving aside the question of the existence and characteristics of causality, the new generation of philosophers tries to construct down-to-earth accounts of causality, especially by referring to practicing scientists' achievements in finding patterns in the empirical data of targeted variables that they collect from experiments or field studies. In other words, contemporary philosophers of causality, recognizing that we human beings are agents of our own knowledge, tend to use their restricted *methodological* lever to tease out indications of the answers of what previously were thought to be questions about metaphysics and epistemology.

Similarly, the conception and application of mechanisms are nothing new in science and philosophy. From the seventeenth century onward, we observe the

development of "mechanical philosophy," represented by the achievements of the giants of science such as Galileo, Descartes, Huygens, Boyle, and Newton. Marie Boas's (1952) seminal article on the establishment of the mechanical philosophy identifies the rise of the mechanical philosophy as due to the development of new the science of mechanism that replaced Aristotelian physics and thus concludes that explanations for the properties of bodies should be based on it (Boas 1952, p. 414). In the contemporary philosophy of science, the first half of the twentieth century also witnessed mechanistic explanations developed in philosophy of science when the discussion was centered on the mechanics and physics (e.g., Nagel 1961). The resurgence of the importance of mechanisms in recent studies in the philosophy of science, however, is not because its application would reduce explanations in other sciences to mechanics and physics (e.g., Nagel's 1961 attempt to reduce biology), but because of its involvement in how scientists actually explain. A number of philosophical characterizations of mechanisms have been recently put forward (Tabery 2004; Skipper and Millstein 2005). Most of them are inspired by biology. Among them, the two most salient accounts are developed by Peter Machamer, Lindley Darden, and Carl Craver (2000)—hereafter, MDC—and by Stuart Glennan (1996, 2002). Glennan's *interactionist* account evolves hand in hand with the literature of causality. His recent definition adopts James Woodward's (2003) interventionist account of causality. In contrast, MDC endeavor to give up causal language entirely.

Briefly, Woodward's view is that "*X* is a total cause of *Y* if and only if under an intervention that changes the value of *X* (with no other intervention occurring) there is an associated change in the value of *Y*" (Woodward 2007, p. 73). More specifically, Woodward clarifies the relationship among the concepts of manipulation, the change-relating property of a relation, and invariance. Inspired by Douglas Gasking's idea that a causal relation is a "means-end" or "producing-by-means-of" relation (Gasking 1955), Woodward refines this causal idea by adding a condition of invariance. According to Gasking, *C* causes *E* in cases in which we can, with the aid of a certain kind of general manipulative technique, produce an antecedent occurrence of kind *C* as a means to bring about a subsequent occurrence of kind *E*. As with Gasking, Woodward agrees that a relation, if it is to be regarded as having causal and explanatory import, must be explicated in terms of manipulation. What is new in Woodward's account is that he further suggests that, for a relation *R* between *C* and *E* to count as being causal and explanatory, relation *R* must be invariant under the manipulation of *C*. That is, the manipulated change in *C* should bring about the change in *E* in the way stated in *R*; otherwise, *C* does not cause *E* in the way stated in *R* and perhaps does not cause *E* at all. Clearly, for Woodward, a causal relation should be a relation that is exploitable by manipulation for the purposes of control. Woodward's account seems to imply that a relation *R* will express a causal relation only if *R* is invariant over a range of interventions.

Accordingly, Glennan (2002, p. S344) recently offered the following definition: "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (Glennan 2002,

p. S344). He thus avoids the notion of laws that was employed in his early studies (e.g., Glennan 1996). In contrast, MDC think such causal language is too vague to characterize the actual specific activities within a mechanism, such as pulling, scraping, or binding. In their *dualist* account, mechanisms are constituted of *entities* and *activities:* "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer et al. 2000, p. 3). Other definitions of mechanisms include one by Bechtel and Abrahamsen (2005) who stress mechanisms as *structures*. They argue that "[a] mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel and Abrahamsen 2005, p. 423). Some philosophers think the differences between these accounts are highly significant, whereas others think they are minor.

Notice that, prior to the emergence of the mechanist approach in philosophy of biology, mechanisms have been investigated by philosophers of social sciences such as Mario Bunge (2004) and Jon Elster (1983, 1998, 2007) and been advocated by economic sociologists Peter Hedström and Richard Swedberg (1998). Like MDC, the advocates of social mechanisms share Francis Crick's view that biologists prefer to think in terms of mechanisms rather than laws (Hedström and Swedberg 1998, p. 3). But it should be noted that, as successfully argued by Carl Craver and Marie Kaiser in their chapter, mechanist philosophers do not deny the epistemic virtue of regularities, as they help scientists search for mechanisms, even as mechanisms in turn help us to understand how regularities and generalizations provide the basis for scientific activities such as explanations, predictions, and control.

A general notion of social mechanisms is aptly characterized by Thomas Schelling, a Nobel Laureate in economics, who defines social mechanisms in contrast with laws, theories, correlations, and black boxes, conceiving them as plausible hypotheses that explain social phenomena, where the explanation is offered in terms of interactions between individuals or between individuals and social aggregates (Schelling 1998). Similarly, Elster (1998) contrasts mechanisms with black boxes (which could provide no explanations) and laws (which provide only deterministic explanations). He defines social mechanisms as "frequently occurring and easily recognizable causal patterns that are triggered under generally unknown conditions or with intermediate consequences" (Elster 1998, p. 45) and regards them intermediates between laws and descriptions (ibid.). Elster's mechanism-based explanations would consist of the form "if conditions C_1, C_2, $\ldots C_n$ obtain, then sometimes E" (Elster 1998, p. 48). At present, the investigations of social mechanisms seem to converge on the accounts developed by philosophers of biology by reevaluating social mechanisms in terms of mechanist philosophy of science (e.g., Hedström and Ylikoski 2010), implying an attempt to reconcile social mechanisms in a broader conception of scientific mechanisms.

3 Biological Causality and Mechanism

Until recently, philosophical accounts of causation and mechanism in the philosophy of biology used a very limited set of philosophical accounts of causation, assuming they used any account of causation at all (e.g., Rosenberg 1985 and Hodge 1987 discuss causation in biology without appeal to any particular account). Perhaps the most common invocation was of Wesley Salmon's "screening off" condition (see, e.g., Brandon 1988; Lloyd 1988; Sober 1984 articulates his own account of causation as an alternative). This is not to say that biologists and philosophers of biology did not appeal to causes—far from it. It seems rather that, until recently, most philosophers of biology did not find accounts of causation such as Lewis's counterfactual account or Salmon's Mark Transmission/Conserved Quantity account particularly useful for illuminating phenomena in biology. Indeed, some recent discussions of causation in the philosophy of biology still do not cite causation literature (e.g., Mitchell and Dietrich 2006); this is not meant as a criticism, but simply to point out, again, that the philosophical literature on causation is sometimes not seen as helpful or necessary for illuminating philosophical issues concerning causation in biology.

However, the recent development of the above-mentioned mechanist and interventionist philosophies has begun to change that. It would be tedious to list all of the philosophers of biology who have drawn on these works, so here is a short sampling: Fehr (2004), Tabery (2004), Reisman and Forber (2005), Waters (2007), Steel (2007), and Illari and Williamson (2010). In trying to understand the explosion of literature on interventionist accounts of causation and mechanist accounts, it is surely no coincidence that whereas accounts such as Salmon's and Lewis's were derived from physics or from traditional philosophical analysis of our everyday language, accounts such as Woodward's and MDC's were derived from the social sciences, economics in particular, and from biology.

With the interventionist account of causation and accounts of mechanisms in ascendancy in the philosophy of biology, an obvious question arises as to the relationship between causation and mechanism in biology. The possible answers to this question, however, are varied and complex. As Roberta Millstein's chapter in this volume notes, philosophers such as Glennan (2009) have argued that there are *two* types of causation, contra the decades of argument over what constitutes *the* account of causation: causal relevance (or causal dependence) and causal production. Woodward's account, falling into the broad category of counterfactual accounts, is supposed to be a causal relevance account, whereas Glennan's, MDC's, and even Salmon's accounts are seen as causal production accounts. If Glennan and others (e.g., Cartwright 2004; Hall 2004) who have argued that there are two types of causation are right, then different philosophers use different accounts of causation because they pick out different phenomena in the world (or different aspects of the same phenomenon, though Millstein argues that Glennan has failed to make his case for natural selection). However, aside from reservations that one might have about there being two accounts of causation, the relationship

between the two accounts is unclear, because both Craver and Glennan incorporate Woodward's views into their accounts of mechanism (Glennan 2002; Craver 2007) and because Woodward himself has described how interventionist accounts of causation can be used as an account of mechanisms. Moreover, Glennan has argued (1996, 2010) that mechanisms can serve as the basis for a theory of causation. So, perhaps causal relevance and causal production (including mechanisms) are tightly linked. Lindley Darden's chapter usefully explores some of the ways in which causes might manifest themselves in mechanisms: activities of entities, stages of mechanisms, or as start or setup conditions. On Darden's view, then, analyses of "mechanism produces phenomenon" are much more detailed and specific than "*C* causes *E*," as the former incorporates many of the latter, plus other aspects such as the ways in which entities and activities are organized.

A second sort of question arises as to which biological phenomena can be profitably illuminated by accounts of causation and/or mechanisms; each of the biology papers in this volume contributes partial answers to this question by exploring causation and mechanism in different areas of biology. Once again, however, we quickly realize that for every illumination, new questions are uncovered. Several of the papers deal with causation and/or mechanisms in evolutionary biology. Millstein, who has elsewhere (2006) argued that natural selection is a population-level causal process, argues (contra Glennan) that the causation at the population level exhibits causal production (in Salmon's sense) as well as causal relevance (a point on which she and Glennan agree). But she does not take a stand on whether natural selection should be understood as a mechanism, having elsewhere (Skipper and Millstein 2005) raised concerns for such a claim. However, Rong-Lin Wang offers some criticisms of Millstein's claim that natural selection is a population-level causal process. For example, he argues that Millstein's account of natural selection does not handle cases of what Elliott Sober has called "selection of" (as distinguished from "selection for" and random drift). Moreover, he suggests that prospects of the view that natural selection is a population-level causal process depend on a satisfactory solution to each of the three problems: the redundant cause problem, the overdetermination problem, and the epiphenomenon problem. Thus, according to Wang, we need to pay attention to the work of metaphysicians in order to understand the nature of selection. The other philosophy of biology papers, discussed elsewhere in this Introduction, explore causation and mechanisms in other aspects of evolutionary biology as well as other areas of biology such as genetics, plant breeding, and biomedicine.

4 Economic Mechanism and Causality

A mechanism is often conceived as a machine, which is on the top of Craver and Darden's (2005) list of ideas associated with the term mechanism. It is so because machines provide *models* of intelligibility that have contributed to our understanding of the mechanisms in the natural world. This understanding of mechanism in

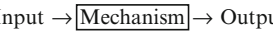


Fig. 1.1 The Bungean input-output mechanism (Adopted from Bunge 2004)

terms of such an artifact is commonplace in both natural and social sciences, since such visualization helps understand the various aspects of mechanisms and how they are constituted. A mousetrap, for instance, is used by Craver and Bechtel (2006) to illustrate the philosophical notion of mechanism. In social science, underneath different definitions of social mechanisms we previously discussed is a strong perception of the society as a machine. An example is Elster's assertion that an explanation in social science requires of specifying *social cogs and wheels*. According to Elster (1983, p. 24), "To explain is to provide a mechanism, to open up the black box and show the nuts and bolts, the cogs and wheels of the internal machinery." For Bunge, a mechanism is specifically perceived as an input-output machine, which is contrasted with black boxes by which inputs and output are connected without knowing the inner machinery (Bunge 2004; Hedström and Swedberg 1998), and can be understood in terms of the diagram in Fig. 1.1.

In economics, economists have also been using and adopting the concept of mechanism for centuries. As Harro Maas (2005) demonstrates, the practices of mechanical reasoning among the political economists had been observed in Victorian Britain, where the economic world—perhaps analogous to the physical world—was envisaged as a machine. More specifically, the type of machine is an *input-output* machine, coinciding with the Bungean input-output mechanism. There are two salient cases in the economic literature. First, the "market mechanism," which is perhaps the most fundamental concept indicating the structure and capacity of the market for allocating the resources among economic units, is commonly understood as such. To study economic mechanisms, a subfield *mechanism design* emerged in the 1960s to apply proper mathematical tools to construct and analyze how economic units and activities are coordinated and guided through the information they receive from a fictitious "information center." A mechanism is thus also viewed as a communication or a dialogue between the information center and economic units or "periphery" (Hurwicz 1973, pp. 6–7). In this regard, economic mechanisms bear some similarity to mental mechanisms as they both are information-processing mechanisms (c.f., Bechtel 2008). Second, the *input-output analysis*, which was established in the 1930 by the economist Wassily Leontief, explicitly treats the structure of an economy as constituted by input-output relationships. With its ambition to quantitatively deal with all components of the economy, the input-output analysis requires its models to be computable and statistically measurable so that it can describe and interpret the economic operations in terms of "directly observable basic structural relationships" (Leontief 1987, p. 860).

However, as of now, it seems causality rather than mechanism is economists' primary concern. Mechanism in general is understood in the context of "causal mechanism," whose structure—causal structure—needs to be identified. Kevin Hoover (2001, p. 24) offers one definition of causal structure as "a network of

counterfactual relations that maps out the underlying mechanisms through which one thing is used to control or manipulate another." While mechanism is defined freely in this definition, the general idea of mechanisms developed in various mechanist accounts can surely apply to it.

With respect to Hoover's description of causal structure, each causal path between any two variables within a causal structure is represented as an invariant counterfactual conditional relation. It is called "counterfactual" because it claims that if there is a "hypothetical" change in (or manipulation of) the supposed causal variable, then the supposed effect variable will have a corresponding degree of change. If we represent the causal relation between two variables p and q as the equation $q = \alpha p + \varepsilon$, where the parameter α represents the degree of change of p in q, then the adjective "invariant" means that, against the background of a complicated network of the causal structure, whatever unit of change in p there is, the corresponding effect of α degree of change of p in q will "remain unchanged." In that case, the fact of invariance can be used as a criterion, as was pointed out by Herbert A. Simon in his 1953 article, that would permit us to discriminate among competing structural representations that are consistent with the same set of data. Based on this view, it is no wonder that Hoover remarks that "causal structure is characterized by a parameterization that governs the manner in which variables are related to each other... The patterns of relative independence, dependence, and interdependence among variables—the causal structure—are dictated by the parameterization" (Hoover 2001, p. 59). Hoover's structural account of causality can be regarded as a classic metatheoretical account that aims to characterize scientists' attempt to use their limited methodological lever—such as the available statistical techniques—to tease out, from the probabilistic distribution of those relevant variables, the indications of the answers of causal inquiries. Hoover's chapter in this volume goes further to explicate the structural approach by contrasting with Woodward's manipulability account, arguing that *modularity*—a critical characteristic of Woodward's account indicating that each equation in a system of causal relations corresponds to a distinct causal mechanism—fails in certain cases, because in reality individual equations in a causal structure do not necessarily correspond to distinct mechanisms. Furthermore, Hoover argues that, unlike Woodward's manipulability account in which the notion of causality is defined in token level (causal relations hold among particular events), the structural account explains causal notion in type level (causal relations hold among variables) and could be more explanatory for causal relationships in a practical sense.

5 Representing Causal Structures and Mechanisms

Given the importance of understanding both causal structures and mechanisms, there is a need for inquiring into the possibilities of providing epistemological access and representation to them. One pivotal question concerns whether we can completely know causal structures and mechanisms. Recall that the notion of

mechanisms is employed by philosophers and scientists in a sense to contrast with black boxes. But whether mechanisms can be completely known remains under debate. For instance, in the above-mentioned economic approaches of mechanism design and input-output analysis, the mechanism of the economy is regarded as being perceivable, given suitable tools—mathematical methods for mechanism design and statistical analysis for the input-output analysis—exist. As Leontief put it, to understand an economy requires nothing but a "direct structural analysis," like a mechanic looking under the hood (Leontief 1954). Leontief thought not only is such a direct observation possible, but it is the only promising way of understanding the operational characteristics of the economy (Leontief 1954, p. 230). By contrast, Trygve Haavelmo, the pioneer of the probabilistic approach to econometrics, used his famous mechanical analogy to illustrate the methodology of econometric models (Haavelmo 1944, pp. 27–8): The empirical relationship between the amount of throttle and the speed of a car, under uniform circumstances, is regular. Such a relationship is useful for driving a car in a prescribed speed, but is not fundamental. The throttle-speed relationship not only lacks of *autonomy* because it breaks down as the condition changes, but also, the relationship tells us little about how the car works, hence it "leaves the whole inner mechanism of a car in complete mystery" (Haavelmo 1944, 371 p. 27).

Thus, while Haavelmo thought that understanding the inner mechanism is of primary importance, he contrasted with the economists such as Leontief in thinking that a direct observation is impossible. This characterizes the practices of econometricians, who have been trying to use a mix of tools from economics, mathematics, and statistics to analyze empirical data and, in part, concerned whether the *data-generating process*, or *DGP*, that is regarded as being responsible for producing the observed data is real or fictitious and whether it can be fully known.[2] Ontology aside, many have maintained that econometric models do not allow observation of the DGP directly. One can receive only an incomplete image of the underlying structure by inferring from observed data. Because, unlike a mousetrap, scientific mechanisms are usually not available for direct observation, hopes for complete descriptions of mechanisms and/or causal structures would be in vain.

Even so, the incomplete notions of mechanisms and causal structures are still useful for understanding science. In order to represent the underlying mechanism, scientists use what MDC called "mechanism schemata" or "mechanism sketches" as incomplete description. For them, mechanism sketches are black boxes, serving to indicate required future research work in order to establish mechanism schemata. Mechanism schemata, in contrast, contain more, but still incomplete, information and are usually represented by diagrams. Since neither sketches nor schemata are thorough and detailed, to understand mechanisms via sketches and schemata might be related to the "black box inference" in the philosophy of science. Although the term was made famous by Sober (1998) who discussed particularly the linkage

[2] See Chao (2009, esp. Ch. 7) for the philosophical discussion on the DGP.

between causes and effects, it is longstanding in the philosophy of science concerning the structure of scientific theories. Hempel (1966), for instance, when discussing the distinction between observables and unobservables, suggests that in an attempt to explain the performance of a black box which "responds to different kinds of input by specific and complex output" (p. 81), the internal structure of the black box is in principle observable, or can be directly inspected, as long as appropriate instruments are available. Hence "any line drawn to divide them into actual physical objects and fictitious entities would be quite arbitrary" (Hempel 1966, p. 82). Similarly, Hanson (1963) illustrates that, as science progresses, our understanding of phenomena switches from the stage of "black box" to that of "grey box" and finally reaches the stage of "glass box" whence the theory and the phenomena are of the same structure and the equations of the theory can actually "mirror" the processes of the nature (Hanson 1963, p. 38). Hanson's account is shared by the mechanists. MDC, for instance, regard the schemata as essential heuristic devices for discovering mechanisms. By reasoning with a schema, scientists are guided to choose known and proper entities or activities to fill the gap. Afterward, when a schema is instantiated, it provides a mechanistic explanation of the phenomena that the mechanism produces (Machamer et al. 2000, p. 29).

It is thus natural to relate mechanism sketches and schemata to scientific models; both serve inferential and representational devices to understanding science. Recent study (e.g., Morgan and Morrison 1999) emphasizes that models are independent of theory and the world and thus have autonomous power for representing each of them. Literature also shows that the distinction between *models of theories* and *models of data* that was earlier made by Patrick Suppes in his influential article "Models of Data" (Suppes 1962) has proven useful for characterizing scientific modeling processes. Following Suppes and the discussion of empirical models in science and philosophy, Ruey-Lin Chen argues in his chapter that scientific discovery in biology can be explained and instantiated through the models of experimental data. In contrast, Till Grüne-Yanoff's chapter in this volume deals with the issue of representing mechanisms at a theoretical level. He examines evolutionary game theory (EGT), arguing that EGT models employed in biology and economics have different interpretations concerning what causal factors and relations they represent, interpretations that are captured by informal mechanism descriptions rather than by the EGT formalism. An abstract model is qualified as MDC's mechanism sketch; it requires an interpretation of the model to represent a specific mechanism in biology or in economics. In other words, biological or economic mechanism descriptions are of a particular kind: They do not describe the composite parts of a system, but they describe in abstract form the stages through which the mechanism runs. Because it does so in a highly abstract way, many different mechanisms can be subsumed under these descriptions, making them general schemata useful for many scientific purposes.

Hoover's and Steel's chapters demonstrate that representational devices such as models can sometimes play a more active role. They both use directed acyclic diagrams (DAGs) to represent causal relations, which has been a popular representational tool employed by philosophers of causality. Hoover points out that

results in causal analysis may not be independent of the modes of representation, that is, equations or graphs, and clarifies the relationships between graphical and equational representation of causality. Steel goes a step further to include DAGs as a part of the definitions of philosophical notions of extrapolation and integration, implying that theoretical propositions could be entranced by directly consisting of representational tools.

Furthermore, the variations in methodologies could be represented by the change in representational tools, and vice versa. Marcel Boumans's chapter revisits his (Boumans 1999) "recipe-making" account of models.[3] It suggests that mathematics provides the means of molding different ingredients into a new model. In a sense, as Boumans points out, early econometricians such as Jan Tinbergen regarded mathematical forms as determining the economic movement. However, when the focus switched to identify causal structural relationship among variables, the primary concern for econometricians was to seek the model's property of invariance. Since the econometricians then adopted the strategy of relying on theories to do the job, the role of mathematical molding was lost. These works show how thinking about representation and models provides new insights into mechanisms and causal structures.

6 Mediation and Extrapolation

Let us return to our original aspirations to bring together biological and economic mechanisms and causality. The idea promoted in this volume is that studies in the philosophy of science would be enriched by two ways of research. The first is to start with the concepts that scientists use most in their practices. Such investigations provide concrete grounds of scientific methods and activities on which philosophical notions can (and arguably should) be based. Readers can observe that most of our chapters attempt to address simultaneously the notions of causality and mechanisms. Though the notions defined and employed in their work—and the cases they study—do not belong to one single account, the plural meanings of mechanisms and causality clearly show their importance to understanding science. The second way is to conduct interdisciplinary explorations on how the concepts are understood by different groups of scientists and philosophers. In this volume, in addition to Grüne-Yanoff's chapter studying biological and economic game-theoretic models, we have three chapters dealing with comparisons and contrasts of facts and methods between economics and biomedical science. All three chapters start with specific scientific works in economics and biomedical science, then conduct methodological investigations on the case studied. One central common theme, which has been dealt with by the authors of this volume under various topics, such as Darden's *interfield integration* (Darden 2006) and Steel's

[3] The term is coined by Mary Morgan (2008).

extrapolation (Steel 2007), is to investigate whether methods, hypotheses, and facts of one field can be applied to another. David Teira and Julian Reiss's chapter compares research methods of randomization in medical and economic sciences: randomized clinical trials (RCTs) and randomized field experiments (RFEs), respectively. Randomized controlled trials have long been regarded as the gold standard for finding causal relations between interventions and experimental outcomes. One reason, as Teira and Reiss point out, is that they provide *mechanical objectivity*, meaning that randomized trials usually follow rigorous and transparent rules so that the results are immune to the bias of subjective expert judgment. They, however, argue that such objectivity is hard to come by. It is because the participants both RCTs and RFEs could act so strategically to obtain their best interest from the trial experiment that the supposed invariance of the controlled environment breaks. Consequently, it is questionable to infer from the evidence the causal connections between treatments and the results.

Both Hsiang-Ke Chao and Szu-Ting Chen's and Steel's chapters deal with the issue of extrapolation that was conceptualized by Steel (2007), and they both deal with the studies that can be categorized as *freakonomics*—using economic principles to (surprisingly) explain a social phenomenon that was first thought to be out of the realm of economics—popularized by the economist Steve Levitt. Steel uses John Donohue and Levitt's (2001) controversial article of the causal relation between legalized abortion and crime rate in the United States. Because there is no direct evidence that can be used to check whether the hypothesis is correct, social scientists support the US case by analyzing results derived from a survey of a similar case that happened in the Scandinavian and Eastern European areas during some periods in the twentieth century. But can we legitimately use evidence obtained in a different time and a different area to support the local case? The problem of extrapolation is analyzed by applying a mechanism-based approach—what Steel calls "comparative process tracing."

Another case where extrapolation could lead to possible explanation is the "missing women" debate discussed by Chao and Chen, in which a biological explanation—hepatitis B virus infection—for the abnormal inequality of sex ratio at birth in Asia is offered by, extrapolated by, and instantiated by the sampling data in the other area. But the biological explanation is claimed to be rejected by economists who used Taiwanese population-level data. They find empirically that cultural factors such as son preference are the cause for the missing women phenomena. Chao and Chen argue that such empirical study does not necessarily deny the existence of the underlying biological causal path, since what is observed is a net causal result. Taking the net causal result as an evidence of ruling out minor causal paths is equal to treating the underlying mechanism as a nontransparent box. In this regard, extrapolations regarding evidence in different time and space can be seen as complementary rather than substitutive.

This echoes our account of the ontology of mechanisms and causal structure: In search of an explanation for a phenomenon, it is adequate to specify the mechanism or identify the causal structure that underlies it. Science progresses thus from black box to grey box and in turn to transparent box, but not the other way around.

7 Conclusion

We have argued in this introductory chapter that thinking about mechanisms and causality enables us to access scientific practices in biology and economics. Methodological investigations centering on mechanisms and causality provide a meaningful way to understand science. The detailed philosophical analysis of these two conceptions, together with specific biological and economic cases given in the chapters of this volume, is our attempt to mediate between mechanisms and causality and between biology and economics. We look forward to seeing more explorations of these topics in future philosophy of science studies.

References

Bechtel, William. 2008. *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. New York: Routledge.

Bechtel, William, and Adele Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Boas, Marie. 1952. The establishment of the mechanical philosophy. *Osiris* 10: 412–541.

Boumans, Marcel. 1999. Built-in justification. In *Models as mediators: Perspectives on natural and social science*, ed. Mary S. Morgan and Margaret Morrison, 66–96. Cambridge: Cambridge University Press.

Brandon, Robert N. 1988. The levels of selection: A hierarchy of interactors. In *The role of behavior in evolution*, ed. H.C. Plotkin, 51–71. Cambridge, MA: MIT Press.

Bunge, Mario. 2004. How does it work? The search for explanatory mechanisms. *Philosophy of the Social Sciences* 31: 182–210.

Cartwright, Nancy. 1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cartwright, Nancy. 2004. Causation: One word, many things. *Philosophy and Phenomenological Research* 71: 805–819.

Cartwright, Nancy. 2007. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

Chao, Hsiang-Ke. 2009. *Representation and structure in economics: The methodology of econometric models of the consumption function*. London/New York: Routledge.

Craver, Carl F. 2007. *Explaining the brain*. Oxford: Oxford University Press.

Craver, Carl F., and William Bechtel. 2006. Mechanism. In *Philosophy of science: An encyclopedia*, ed. Sahotra Sarkar and Jessica Pfeifer, 469–478. New York: Routledge.

Craver, Carl F., and Lindley Darden. 2005. Introduction. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 233–244.

Darden, Lindley. 2006. *Reasoning in biological discoveries*. New York: Cambridge University Press.

Donohue, John, and Steven Levitt. 2001. The impact of legalized abortion on crime. *Quarterly Journal of Economics* 116: 379–420.

Elster, Jon. 1983. *Explaining technical change: A case study in the philosophy of science*. Cambridge: Cambridge University Press.

Elster, Jon. 1998. A plea for mechanisms. In *Social mechanisms: An analytical approach to social theory*, ed. Peter Hedström and Richard Swedberg, 45–73. Cambridge: Cambridge University Press.

Elster, Jon. 2007. *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge: Cambridge University Press.
Fehr, Carla. 2004. Feminism and science: Mechanism without reductionism. *Feminist Formations* 16: 136–156.
Gasking, Douglas. 1955. Causation and recipes. *Mind* 64: 479–487.
Giere, Ronald N. 1988. *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.
Giere, Ronald N. 1999. *Science without laws*. Chicago: University of Chicago Press.
Giere, Ronald N. 2008. Naturalism. In *The Routledge companion to philosophy of science*, ed. Stathis Psillos and Martin Curd, 213–223. London/New York: Routledge.
Glennan, Stuart. 1996. Mechanisms and the nature of causation. *Erkenntnis* 44: 49–71.
Glennan, Stuart. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69: S342–S353.
Glennan, Stuart. 2009. Productivity, relevance, and natural selection. *Biology and Philosophy* 24: 325–339.
Glennan, Stuart. 2010. Mechanisms, causes, and the layered model of the world. *Philosophy and Phenomenological Research* LXXXI: 362–381.
Haavelmo, Trygve. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement): 1–115.
Hacking, Ian. 1983. *Representing and intervening: Introductory topics in the philosophy of science*. Cambridge: Cambridge University Press.
Hall, Ned. 2004. Two concepts of causation. In *Causation and counterfactuals*, ed. J. Collins, N. Hall, and L.A. Paul, 225–276. Cambridge, MA: The MIT Press.
Hanson, Norwood Russell. 1963. *The concept of the positron: A philosophical analysis*. Cambridge: Cambridge University Press.
Hedström, Peter, and Richard Swedberg. 1998. Social mechanisms: An introductory essay. In *Social mechanisms: An analytical approach to social theory*, ed. Peter Hedström and Richard Swedberg, 1–31. Cambridge: Cambridge University Press.
Hedström, Peter, and Petri Ylikoski. 2010. Causal mechanisms in the social sciences. *Annual Reviews of Sociology* 36: 50–67.
Hempel, Carl G. 1966. *Philosophy of natural science*. Englewood Cliffs: Prentice-Hall.
Hodge, M.J.S. 1987. Natural selection as a causal, empirical, and probabilistic theory. In *The probabilistic revolution, Vol. 2: Ideas in the sciences*, ed. Lorenz Krüger, Gerd Gigerenzer, and Mary S. Morgan, 233–270. Cambridge: MIT Press.
Hoover, Kevin D. 2001. *Causality in macroeconomics*. Cambridge: Cambridge University Press.
Hurwicz, Leonid. 1973. The design of mechanisms for resource allocation. *American Economic Review* 63: 1–30.
Illari, Phyllis McKay, and Jon Williamson. 2010. Function and organization: Comparing the mechanisms of protein synthesis and natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences* 41: 279–291.
Leontief, Wassily. 1954. Mathematics in economics. *Bulletin of the American Mathematical Society* 60: 215–233.
Leontief, Wassily. 1987. Input-output analysis. In *The new Palgrave. A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, 860–864. London: Macmillan.
Lloyd, Elisabeth A. 1988. *The structure and confirmation of evolutionary theory*. Westport: Greenwood Press.
Maas, Harro. 2005. *William Stanley Jevons and the making of modern economics*. Cambridge: Cambridge University Press.
Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67: 1–25.
Millstein, Roberta L. 2006. Natural selection as a population-level causal process. *The British Journal for the Philosophy of Science* 57: 627–653.

Mitchell, Sandra D., and Michael R. Dietrich. 2006. Integration without unification: An argument for pluralism in the biological sciences. *American Naturalist* 168: S73–S79.
Morgan, Mary S. 2008. Models. In *The New Palgrave dictionary of economics*, 2nd ed, ed. Steven N. Durlauf and Lawrence L. Blume. London: Palgrave Macmillan.
Morgan, M.S., and M. Morrison (eds.). 1999. *Models as mediators: Perspectives on natural and social science*. Cambridge: Cambridge University Press.
Nagel, Ernest. 1961. *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt, Brace & World.
Reisman, Kenneth, and Patrick Forber. 2005. Manipulation and the causes of evolution. *Philosophy of Science* 72: 1113–1123.
Rosenberg, Alexander. 1985. *The structure of biological science*. Cambridge: Cambridge University Press.
Schelling, Thomas C. 1998. Social mechanisms and social dynamics. In *Social mechanisms: An analytical approach to social theory*, ed. Peter Hedström and Richard Swedberg, 34–44. Cambridge: Cambridge University Press.
Skipper Jr., Robert A., and Roberta L. Millstein. 2005. Thinking about evolutionary mechanisms: Natural selection. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 327–347.
Sober, Elliot. 1984. *The nature of selection*. Cambridge, MA: MIT Press.
Sober, Elliot. 1998. Black box inference: When should intervening variables be postulated? *The British Journal for the Philosophy of Science* 49: 469–498.
Steel, Daniel. 2007. *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.
Suppes, Patrick. 1962. Models of data. In *Logic, methodology and philosophy of science: Proceedings of the 1960 international congress*, ed. Ernest Nagel, Patrick Suppes, and Alfred Tarski, 252–261. Stanford: Stanford University Press.
Tabery, James. 2004. Synthesizing activities and interactions in the concept of a mechanism. *Philosophy of Science* 71: 1–15.
Van Fraassen, Bas. 1980. *The scientific image*. Oxford: Oxford University Press.
Van Fraassen, Bas. 1989. *Law and symmetry*. Oxford: Oxford University Press.
Waters, C. Kenneth. 2007. Causes that make a difference. *Journal of Philosophy* CIV: 551–579.
Woodward, James. 2003. *Making things happen*. Oxford: Oxford University Press.
Woodward, James. 2007. Causation with a human face. In *Causation, physics, and the constitution of reality: Russell's republic revisited*, ed. H. Price and R. Corry, 66–105. Oxford: Oxford University Press.

Part I
Defining Mechanism and Causality

Chapter 2
Mechanisms Versus Causes in Biology and Medicine

Lindley Darden

Abstract Biologists use knowledge of mechanisms for explanation, prediction, and control. Philosophers of biology, working in the new mechanistic philosophy of science, have identified features of an adequate description of a biological mechanism. The very abstract schema term "cause" may refer to any of various components of a mechanism, or even conditions needed for it to operate. A case study of the disease cystic fibrosis illustrates the advantages (and complexities) of identifying the various stages of the relevant mechanisms. Such knowledge is more useful than merely claiming that a mutation in the CFTR gene causes the disease, given the goals of explanation, prediction, and control of disease symptoms. Knowledge of "mechanism produces phenomenon" is often much more useful for explanation, prediction, and control than "C causes E."

1 Introduction

Contemporary biologists often seek to discover mechanisms. Many such discoveries were major achievements in twentieth-century biology, such as the mechanism of Mendelian heredity (Morgan et al. 1915; Darden 1991), the numerous mechanisms of cellular metabolism (Bechtel 2006), mechanisms in neuroscience (Craver 2007), and the mechanisms of DNA replication, protein synthesis, and gene expression in molecular biology (Watson et al. 2007; Darden and Craver 2002). Philosophers of biology are now studying the nature of biological mechanisms in "the new mechanistic philosophy" (Skipper and Millstein 2005).

The team of Peter Machamer, Lindley Darden, and Carl Craver characterized mechanisms and applied that characterization to cases from molecular biology and

L. Darden (✉)
Department of Philosophy, University of Maryland, 1125A Skinner Building, College Park 20742, MD, USA
e-mail: darden@umd.edu

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_2,

neurobiology (Machamer et al. 2000; hereafter referred to as MDC). Others worked on mechanisms in such fields as biochemistry and cell biology (Bechtel and Richardson 1993, 2010; Bechtel 2006), evolutionary theory (Barros 2008), medicine (Thagard 1998; Moghaddam-Taaheri 2011), and the social sciences (e.g., Hedström 2005). Philosophers work to analyze the relation of this new work on mechanisms to traditional topics in philosophy of science, such as explanation (Bechtel and Abrahamsen 2005; Craver 2007) and causation, addressed in diverse ways by Jim Bogen (e.g., 2004, 2005, 2008), Bill Bechtel and Carl Craver (e.g., Craver and Bechtel 2007; Craver 2007, Ch. 3), Stuart Glennan (1996, 2002, 2010), and Jim Woodward (e.g., 2002).

Biologists seek mechanisms for three reasons: explanation, prediction, and control. In this chapter, I will argue that within the mechanistic sciences, such as molecular biology and molecular medicine, the claim "C causes E" is impoverished compared to the claim that "this mechanism produces this phenomenon." Knowledge of a mechanism in the biological sciences is usually more useful for explanation, prediction, and control than merely being able to label something as a cause. Furthermore, the new mechanists emphasize the importance of characterizing (and recharacterizing as work proceeds) the phenomenon that the mechanism produces. Such characterization is a rich description, providing guidance and constraints in the search for the mechanism.

I proceed as follows. In Sect. 2, I summarize one current view of biological mechanisms, the MDC characterization of biological mechanisms. In Sect. 3, I first summarize what we said in the MDC paper about the relation of the analysis of mechanism to an analysis of cause. Then, I expand it to conjecture what "C causes E" might refer to, from the perspective of biological mechanisms. In Sect. 4, I take up the extension of the MDC account to medicine and illustrate the power and complexities that the search for mechanisms plays in an example from medicine. Medical researchers seek mechanisms not just to give explanations for disease symptoms but also to predict the occurrence and severity of the disease and control the outcome for the patient's benefit. We might say: "A mutation in the CFTR gene causes cystic fibrosis." But that is much too simple. To illustrate the usefulness of knowledge of mechanisms, I trace the history of our understanding of the mechanisms that account for, and therapies to treat, the disease of cystic fibrosis. This example illustrates general features about the role of discovering mechanisms for explanation, prediction, and control in fields with practical aims, such as medical research.

2 The MDC Characterization of Mechanisms

A mechanism is sought to explain how a phenomenon is produced. Our team of Machamer, Darden, and Craver characterized mechanisms in the following way:

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (MDC 2000, p. 3)

The MDC *characterization* of mechanisms is not a *definition* giving necessary and sufficient conditions for the term's usage in all cases. Instead it is a characterization to capture the way biologists use the term, as informed by our detailed examination of cases from molecular biology and neurobiology and also informed by philosophical reflection on requirements for *productive* changes.

An example of a biological mechanism is the mechanism of protein synthesis. From the beginning of the field of molecular biology in the 1950s, one of the phenomena puzzling biologists was how proteins are synthesized. By the 1970s, molecular biologists and biochemists had discovered the key details of the mechanism of protein synthesis (Darden and Craver 2002). The mechanism is often represented by the abstract schema, called the "central dogma" of molecular biology:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

It may also be represented by much more detail as in Fig. 2.1, with structures of entities, the organization of the mechanism components within a cell, and the temporal stages and movements depicted by arrows. The mechanism begins in the nucleus with the unwinding of the DNA double helix and the synthesis of messenger RNA. The long ribbon of mRNA moves into the cytoplasm where it attaches to the cell organelle, the ribosome. The ribosome is the site where transfer RNAs, carrying their respective amino acids, attach to the messenger RNA (in a specific order, determined by the genetic code). The growing chain of amino acids will later leave the ribosome and fold into a three-dimensional protein (not shown in Fig. 2.1).

This example illustrates many of the general features of biological mechanisms. These are listed in Table 2.1. The first feature is "phenomenon" because the first step in the search for a mechanism is to identify and characterize a puzzling phenomenon of interest. Next are componency features. The mechanism is composed of entities and activities, sometimes further organized into functional modules. Functional modules are groups of entities and activities that play a given role in the mechanism and may recur in mechanisms of the same abstract type, e.g., the module for translation in the mechanism of protein synthesis (discussed below).

Note that the entities in the protein synthesis mechanism are not all at the same size level. Working entities of the protein synthesis mechanism range from small ions to larger macromolecules to cell organelles (composed of macromolecules). Size level and mechanism level need not, and often do not, correspond (Craver 2007, Ch. 5). Mechanisms have working components of a certain size, with structure and with other properties that enable them to engage in the activities that drive the mechanism.

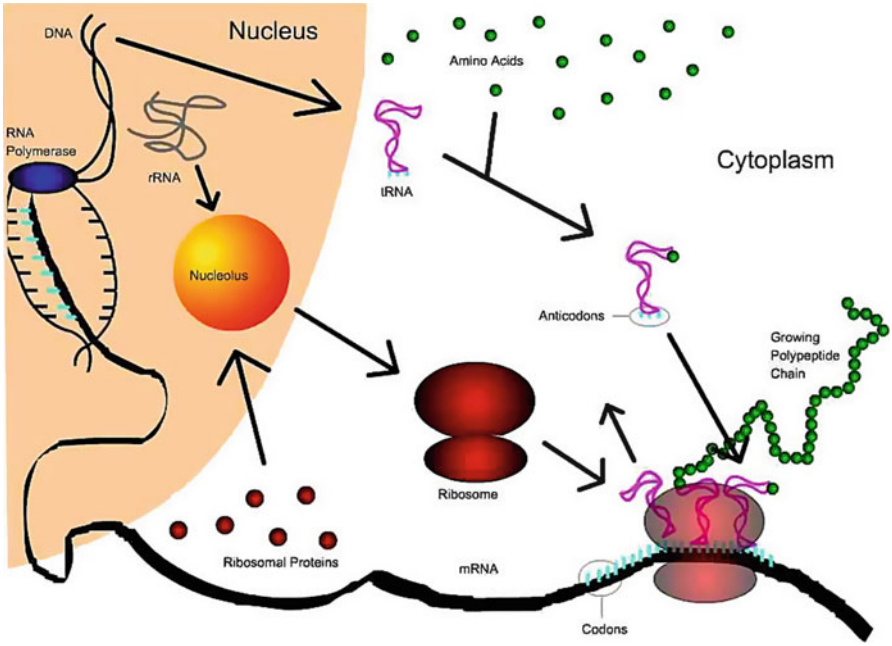


Fig. 2.1 Mechanism of protein synthesis

Table 2.1 Features of mechanisms

Phenomenon
Components
Entities and activities
Modules
Spatial arrangement of components
Localization
Structure
Orientation
Connectivity
Compartmentalization
Temporal aspects of components
Order
Rate
Duration
Frequency
Contextual locations
Location within a hierarchy
Location within a series

Modified from Darden (2006, Table 12.1) and Craver and Darden (2001, Table 2.1)

The mechanism's components have spatial and temporal organization. Spatial organization includes location, internal structure, orientation, and connectivity (both among component parts within the mechanism and to other mechanisms before the start condition and after the termination condition). Sometimes a molecular mechanism is compartmentalized, e.g., occurring in one part of the cell and surrounded by a membrane that protects its parts from dissipation and attack or from attacking other parts of the cell. (Lysosomes, e.g., contain caustic enzymes that break down waste materials; their enzymes are enclosed in that cell organelle and thus do not attack other cellular components.) Also, the stages of the mechanism occur in a particular order and they take certain amounts of time (duration). Some stages occur at a certain rate or repeat with a given frequency. In addition to the componency, spatial, and temporal features of a mechanism, the mechanism may be situated in wider contexts—in a hierarchy of mechanism levels (Craver 2007, Ch. 5) and in a temporal series of mechanisms (Darden 2005). These features of mechanisms can play roles in the search for mechanisms, and then they become parts of an adequate description of a mechanism. What counts as an adequate description (i.e., how much detail needs to be specified) depends on the context in which an explanation of the puzzling phenomenon is sought and the purposes for which the description is to be used.

One use is to make predictions. When the mechanism is in place and the start conditions obtain, then the orderly operation of each stage of the mechanism results in the production of the phenomenon. Hence one can predict what the outcome will be. However, if a portion of the mechanism is broken, then one can predict that the earlier stages operated and an intermediate product accumulates (or perhaps no product at all is produced). Knowing about the intermediate stages allows more fine-grained predictions about what is the output of each stage and what will happen when a stage breaks. A scientist may be able to run a mental simulation of the mechanism and thereby predict what phenomenon it will produce or to predict what will happen if a part of the mechanism is broken. (On mental simulations of mechanisms operating, see Bechtel and Abrahamsen 2005.) However, sometimes the complexity of the mechanism makes mental simulations difficult. Computational simulation models of the mechanism are more useful, especially for quantitative predictions about, e.g., concentrations of products (e.g., Eisenhaber 2006) or for predicting complex spatial interactions as in molecular dynamic simulations (e.g., Watanabe et al. 2010).

A *mechanism schema* is a truncated abstract description of a mechanism that we know how to fill with more specific descriptions of component entities and activities, such as the schema for the central dogma, discussed above. In contrast, a *mechanism sketch* cannot (yet) be instantiated. Components are (as yet) unknown. Sketches may have black boxes for missing components that are sought as the search for the mechanism proceeds. An adequate description of a mechanism (in the context of a given puzzling phenomenon) is an account with all the black boxes filled, with the overall organization specified (e.g., linear or cyclic), and with the features of Table 2.1 noted.

3 Mechanisms and Causes

In this section, I briefly discuss ways that talk of "cause" and "effect" may possibly be mapped to talk of "mechanism" and "phenomenon." This is not a thorough discussion of the many topics addressed by those analyzing causation. Rather it is just a brief foray, from the perspective of some of the recent work on biological mechanisms, to show how much more impoverished talk of "causes" is compared to talk of "mechanisms."

Possible referents of the term "cause" are many and varied from the mechanistic perspective. Something that is designated as a cause may refer to a piece of a mechanism. MDC analyzed mechanisms as composed of both entities (with their properties) and activities. Activities are producers of change; they are constitutive of the transformations that yield new states of affairs. As Machamer (2004) noted, activities are often referred to by verbs or verb forms (e.g., participles, gerunds). Molecules *bond*, helices *unwind*, ion channels *open*, and chromosomes *pair* and *separate*.

In MDC, we discussed the relation between cause and activity:

> Activities are types of causes. Terms like "cause" and "interact" are abstract terms that need to be specified with a type of activity and are often so specified in typical scientific discourse. (MDC 2000, p. 6)

We followed Elizabeth Anscombe (1971, p. l37), who noted that the word "cause" itself is highly general. It needs to be specified by other, more specific, causal verbs. Anscombe included the following in her list: scrape, push, dry, carry, eat, burn, and knock over.

Activities are one way to specify causes. An important feature of activities is that they come in types that have been discovered as science has changed. Over the centuries, scientists discovered new types of activities and their ways of operating. Once they are discovered and their modes of operating well understood, types of activities become part of the "store" or "library" of mechanism parts used to construct mechanistic hypotheses in a particular biological field (Darden 2001; Craver and Darden 2001). The kinds of activities most important in molecular biological mechanisms are, first, the push/pull geometrico-mechanical activities familiar since the beginning of the seventeenth-century mechanical worldview and, second, the many forms of chemical bonding discovered in the nineteenth and early twentieth centuries. Each field finds the activities that drive its mechanisms. A major advantage of the MDC view of causes as types of activities is that the vague term "cause" must be made more specific. The specific way that a specific change is brought about must be found in order to have an adequate description of a mechanism.

Methodologically, activities can sometimes be identified independently of the specific entities that engage in them. For example, the melting temperature of the DNA double helix indicated that it contained weak hydrogen bonds, even before the specific subcomponents (the DNA bases) exhibiting those bonds had been identified. More generally, activities may sometimes be investigated to find their

order, rate, duration, and sphere of influence more or less independently of the entities that engage in those activities.

A specific kind of activity produces a specific kind of change. Finding necessary and sufficient conditions to characterize the many diverse kinds of production is difficult and not required for their scientific discovery (Bogen 2008). Rather than seeking a general definition of production, it is more insightful to consider specific kinds of activities and the means for discovering them. As Machamer suggested in MDC, human beings directly experience many kinds of activities, such as collision, pushing, pulling, and rotating—the activities in mechanisms often discussed in the seventeenth century. Scientists have since discovered many kinds of activities not directly detectable by human senses, such as attraction and repulsion, electromagnetism, and movements across membranes to achieve equilibrium. Science students must be trained to understand how these activities work so that, with education, they can "see" (understand) how mechanisms employing them operate.

Moving beyond what MDC claimed about activities and causes, I note that relating "C causes E" to mechanisms may call attention to some piece of a mechanism other than an activity. As Stuart Glennan (1996) notes, analysis of "C causes E" may require an entire underlying mechanism to lay out all the stages between C and E. In such a case, C refers to the entire mechanism at a lower mechanism level. Alternatively, C may refer to an early stage of the mechanism (consisting of entities and their activities) with the other stages between C and E left unspecified. Hence, "cause" may refer to nearer or more distant stages in the mechanism, prior to the stage (E) of interest.

In addition to entities and activities and organization, MDC noted that mechanisms have "start or setup conditions." If a mechanism requires a signal or start condition (some don't, e.g., some biological mechanisms run continuously), then that may be called a "triggering cause" or a "sufficient cause." When the trigger is present (and the set conditions are available), the mechanism begins to operate. Something called a "necessary cause" might be any nonredundant part of the mechanism or, instead, part of the setup conditions for a mechanism to operate. Setup conditions for mechanisms are many and varied. Although some of the setup conditions are known and indicated (such as in the materials and methods sections of scientific papers), they cannot be fully specified, even in controlled laboratory conditions. (This issue is well known in discussions of ceteris paribus conditions.)

My goal in this section thus far has been to try to map the C of "C causes E" onto some piece of a mechanism or to its start or setup conditions. Now let's turn to E, the effect. Presumably that corresponds to the phenomenon of interest. An important starting point for finding a mechanism is to characterize the puzzling phenomenon that the mechanism produces (on recharacterizing the phenomenon as research on the mechanism proceeds, see Bechtel and Richardson 2010). Presumably, the characterization of the effect is similarly important in constraining adequate claims about its cause.

One of the aims in finding causes is to enable humans to exert control. As is sometimes said, a cause is a handle that can be turned to do something. What we wish to control is E, the outcome. The goal of control of the outcome is especially

important in medicine, so now we turn to an example from that field. The example shows that knowledge of the mechanisms operating or failing to operate provides a better handle than knowing that a single "X causes disease Y."

4 Control of the Outcome in the Disease of Cystic Fibrosis

One might say: "A mutation in the CFTR gene causes the disease cystic fibrosis." But this is an impoverished claim, compared to a description of the myriad mechanisms involved in the etiology of the disease.

In medical contexts, the puzzling phenomenon may be described and redescribed in various ways, as work proceeds to discover the mechanisms producing the disease. Also, the phenomenon of interest is different for those attending to different stages in the progress of the disease. Yet other characterizations of the phenomenon may be provided by those doing fundamental research versus those tasked with treating patients, so this case also illustrates different ways of characterizing the phenomenon within the contexts of pure versus applied research. Different characterizations of the phenomenon focus attention on different mechanisms or stages of a given mechanism involved in the disease etiology.

The phenomenon to be explained in what has come to be called the disease of "cystic fibrosis" changed over time, as groups of symptoms were clustered, the gene discovered, some of the activities of the malfunctioning protein found, and later stages of the disease dissected. One can tell a tidy story about the discovery of the normal mechanism, about the many ways it can break, and about how this knowledge has been and is being used in designing drug therapy. This perspective views disease as a broken-normal mechanism and therapy as aimed at restoring normal functioning (Moghaddam-Taaheri 2011).

However, one can view the medical mechanistic picture in a more complex way. One can ask: Is there some other mechanism that can restore chloride transport function rather than fixing the broken mechanism? Alternatively, as is common in medicine, one can just focus on mechanisms that will aid in alleviation of symptoms of those living with the disease. More specifically, one can seek drugs that will aid in preventing the lung infections that typically lead to death for cystic fibrosis patients. For some of these cases, the current understanding of the mechanisms provides powerful tools for medical researchers, but for other cases many black boxes remain.

In the early 1990s, it looked as if the story of conquering cystic fibrosis would be a simple one: gene discovered, mechanism and mutations understood, and guidance provided for therapies for intervention. However, the genotype-phenotype relations are more complex than anyone studying a disease (seemingly) produced by a single gene defect had reason to expect. Some aspects of the connections between the gene mutations and protein defects and the many phenotypic symptoms of the disease are still not well understood.

The following subsections proceed as follows. First, I recount the history of the discovery of the gene associated with cystic fibrosis. Then I describe the different mechanisms associated with the disease of cystic fibrosis, based on different ways of choosing the puzzling phenomenon of interest. The characterization of the phenomenon is a crucial step in delineating the mechanism of interest. The choice of the phenomenon (the effect?) and the goal of the research focus attention on different aspects of a single mechanism (the cause?) or on different mechanisms (different causes?) within the framework of a single disease. It is much too simple to say that a single mutation in a single gene causes the disease of cystic fibrosis.

4.1 History of Cystic Fibrosis Prior to the Discovery of the Gene

Work in the early to mid-twentieth century connected symptoms in the lungs, pancreas, and sweat glands. Medical researchers found that recurrent respiratory infections, raised levels of chloride in sweat, and insufficient pancreatic enzymes were all problems in the epithelial tissues in those organs and glands. The disease was named "cystic fibrosis," but the specific nature of the defect in epithelial tissues was unknown until the 1980s (Knol 1995).

4.2 Discovery of the CFTR *Gene*

Mitchell Drumm (2001), a graduate student and then postdoc who worked in Francis Collins's lab at the University of Michigan in the 1980s, wrote a lively first person account of the discovery of the gene involved in cystic fibrosis (CF). When these medical researchers started their investigation of CF, all the aspects of the molecular genetic mechanism were a black box. Population genetic studies of families with CF patients had shown that the disease is hereditary, not sex linked, and requires two copies of the mutant gene to produce the disease symptoms; carriers with one copy are not sick. In more technical genetic terms, it is an autosomal recessive disease. It is more prevalent in those with Caucasian European ancestry than among other groups in the USA. Before 1989, the gene was not known and the protein it produces was unidentified. However, earlier work on the symptom of salty sweat indicated that the protein was involved in the transport of chloride in and out of the cell (Quinton 1983; discussed in Pearson 2009).

By the 1980s, molecular biological techniques for finding a gene could proceed quickly if the protein and its accompanying messenger RNA could be identified. A complementary DNA, called "cDNA," could be constructed from the messenger and then used as a probe for finding the nuclear DNA and the location on the chromosome where the gene resided. But the search for the CF gene had to proceed without such technological reversal of those later stages of the mechanism. It was

the first gene to be discovered whose protein product was not known beforehand (Drumm 2001, p. 86).

Three groups in North America collaborated in the gene's discovery, bringing different techniques and areas of expertise. Lap-Chee Tsui at the Hospital for Sick Children in Toronto screened the chromosomes of families with CF children, locating the gene on chromosome 7, near certain known markers. Francis Collins's lab at the University of Michigan did the molecular analysis of the chromosome by a process that Collins had invented, called "chromosome jumping." The DNA of the chromosome was chopped up and circularized. Using this chromosome jumping technique, the Collins lab group found related markers more quickly than permitted by the slower technique called "chromosome walking," which required more laborious analysis of linear sequence overlaps. The third collaborator was John Riordan, also in Toronto in the 1980s, who constructed complementary DNA libraries, using messenger RNA from CF tissues. Putative stretches of DNA could be matched against the cDNAs to see if that gene was active in CF tissues.

A comparison between a putative normal gene and the same stretch of DNA from a CF patient found that three bases were missing in the disease gene. As Drumm remarked: "I think we were all expecting a more striking change in the gene if it were truly a mutation that caused CF" (Drumm 2001, p. 87). The gene was sequenced and various hypotheses proposed as to its functional role in cellular mechanisms. (On functions from a mechanistic perspective, see Craver 2001.) Given the similarity of some of its structural domains to other sequences whose function was known, the protein looked like it would reside in the cell membrane and conduct chloride ions across the membrane. Collins, Tsui, and Riordan named it the "cystic fibrosis transmembrane conductance regulator"—"CFTR" for short—in three papers published in *Science* in 1989 (Kerem et al. 1989; Riordan et al. 1989; Rommens et al. 1989).

The CFTR gene is large, with approximately 180,000 base pairs on the long arm of chromosome 7. It produces a large protein with 1,480 amino acids, organized into several different functional domains. Several classes of mutations produce the disease. Researchers have identified the specific locations of the mutations within the gene and traced the different ways each mutant breaks the mechanism. Some mutations are so severe that no protein is synthesized. However, the mutation that occurs in about 90 % of patients with cystic fibrosis in the USA (Rowe et al. 2005) is less severe. Three bases are deleted in the CFTR gene. During protein synthesis, this deletion results in one missing amino acid: phenylalanine at position 508 (of the 1,480 amino acids). Although missing only one amino acid, such Delta F 508 mutant proteins do not fold properly. The misfolded proteins do not implant into the cell membrane to properly transport chloride ions in and out of the cell (Kirk and Dawson 2003). Normally, the cellular machinery degrades misfolded proteins, but not all such mutant protein is degraded (important in potential drug therapy as we will discuss below). Details about the mechanism of degradation, or lack thereof, are black boxes (Bridges 2003).

4.3 Mechanisms Related to Cystic Fibrosis

So, there is a tidy story that we can now tell about the normal gene and the synthesis of the normal CFTR protein and about how different mutations produce different defects. Consider the mechanism for producing the protein with the Delta F 508 mutation. Each stage of the mechanism becomes a potential target for therapy. As Susan Lindee has discussed, the early hope was for gene therapy to replace the defective gene. The many problems with this approach include finding an appropriate vector for delivering the large gene, getting the gene into the appropriate cells (even in the lung cells which are more accessible than those in other organs), getting the gene to a safe location (either chromosomal or an extrachromosomal plasmid) so as not to disrupt other mechanisms, getting sufficient amounts of genes into the cells, preventing the immune system from rejecting any foreign matter used to take the gene into the cell, and getting the genes to respond to cellular regulatory signals to turn on the gene but not to overproduce the protein (Curlee and Sorscher 2003). These problems have yet to be solved; the prospects for successful gene therapy look dim in the case of CF (Lindee and Mueller 2011).

So, consider the next module of the mechanism, the one after the gene itself, as a target: the messenger RNA. The CFTR gene contains not only the coding sequences that eventually direct the ordering of amino acids during protein synthesis but also spacer segments, called introns. A cell organelle, called a "spliceosome," processes the pre-mRNA to produce the mRNA; the spliceosome accomplishes this by snipping out the introns and binding the remaining coding segments together into the final messenger RNA. Researches have succeeded in inserting a minigene into the DNA of human lung tissue grafted onto a mouse. The minigene has the correct coding segment rather than the Delta F 508 three-base mutation. The gene is expressed at the same time as the CFTR gene, thereby overcoming one of the barriers to gene therapy. Then the splicing machinery is induced to put the correct segment into the processed messenger RNA rather than the mutant segment. Some success in the mouse system makes this look promising. However, it is still a long way from human clinical trials (Liu et al. 2002, 2005; discussed in Thomson 2002).

Currently, a primary area for targeted drug therapies is the next stage of the mechanism: the synthesis of the misfolded protein. For the Delta F 508 mutant, the three missing bases in the gene result in one amino acid missing from the protein, which then misfolds. Although some of the protein degrades, some of the misfolded protein remains in the cells. Therapy can be directed to finding drugs that aid in rescuing the undegraded misfolded protein so that it refolds and inserts into the cell membrane and functions (albeit at a reduced level) to transport chloride ions. A robotic process has screened millions of compounds for their effects on the misfolded protein and some promising drug candidates have been found. One is curcumin, a major constituent of the spice turmeric, which has shown promising effects in vitro and in mice models (Rowe et al. 2005, p. 1999).

In contrast to this random screening, rational drug therapy is also being explored. Medical researchers are using a more detailed understanding of the role of

additional molecules to try to correct the defect. These additional molecules, called chaperones, aid the CFTR protein to fold properly (Wang et al. 2006). The discovery of the role played by such additional molecules that interact with CFTR (produced by additional genes, called "modifier" genes) may explain a puzzling phenomenon about the relation between genotype and phenotype. It is puzzling why patients with the same two Delta F508 mutations can still vary in the severity of symptoms of the disease. One hypothesis is that this difference is due to different modifier genes in their DNA. Although cystic fibrosis seemed to be an ideal case of a disease caused by a mutation in a single gene, we can no longer hold such an overly simple view. The mechanism by which modifier genes work becomes important also.

Thus, we see the importance of the way the puzzling phenomenon is characterized in order to focus attention on the relevant aspect of the mechanism. When the puzzling phenomenon is the synthesis of the normal CFTR protein, that mechanism is fairly well understood. But when the puzzling phenomenon is why the Delta F 508 mutant protein fails to function properly, aspects of the mechanism by which the mutant form of the protein is synthesized and misfolded and degraded still have black boxes. Nonetheless, enough is known about that module of the mechanism to guide drug discovery efforts to find drugs to aid with refolding the misfolded protein.

However, when the puzzling phenomenon is a broader one, namely, how the mutant in the CFTR gene produces the symptoms of cystic fibrosis disease in the myriad organs that it affects, many of the details of these mechanisms are unknown. When what is taken to be puzzling is much later in the progress of the disease, even more black boxes remain. What are the stages of the mechanism leading to the thick airway mucus in the lungs that result in the fact that, as cystic fibrosis patients age, they become more susceptible to particular strains of bacteria that are more resistant to treatment? Various hypotheses as to how to fill this black box abound. As a recent review article said: "So far, a unifying mechanism responsible for the vast clinical expression of the disease in the CF airway has not been identified" (Chmiel and Davis 2003, p. 173).

There are even competing hypotheses, which may not be mutually exclusive, about why the airway mucus is thick and particularly susceptible to bacterial infections. Several hypotheses depend on the effects of malfunctioning chloride transport, leading to an imbalance of salt homeostasis or abnormal water absorption producing thicker mucus (Widdicombe 2003). However, new evidence points to a malfunctioning immune response. Neutrophils, which are a type of white blood cell, are recruited to fight bacteria. CF patients also have defective regulation of neutrophils, leading to an overabundance of them. The mechanism for this malfunctioning regulation of neutrophils is not well understood, although some of the entities and their activities have been identified (Gu et al. 2009). As neutrophils break down, the debris, especially their DNA, accumulates in thick mucus that is a site for colonization preferred by certain forms of bacteria. So, if the phenomenon that the physician wishes to alter is the overgrowth of specific strains of bacteria, then the therapeutic effects may be directed to neutrophil regulation, a much later

stage in the disease with different targets than the CFTR protein biosynthesis mechanism (Chmiel and Davis 2003).

This hypothesized mechanism of overexpression of the immune response to inflammation led to the unexpected prediction that anti-inflammatory drugs would be beneficial to CF patients. Without this hypothesis, one would have expected that anti-inflammatory drugs, such as ibuprofen, would have deleterious effects for lungs susceptible to infections. The normal inflammatory response, which recruits neutrophils to the site of an infection, is beneficial in the fight against bacteria. However, because the hypothesized mechanism suggested overexpression of this response, the drug therapy to reduce the response was subjected to a clinical trial, with some success (Konstan et al. 1995).

So, this case shows the many different ways the puzzling phenomenon can be identified and consequently the many different mechanisms that provide candidate "causes" for that chosen phenomenon. If the phenomenon to be explained is the synthesis of the normal CFTR protein, then the mechanism for that is well understood. If the question is the following—"what is the nature of the failure in that gene that leads to cystic fibrosis disease?"—then the answer is that there are many mutations that disrupt that mechanism in different ways (as we discussed, different classes of mutants disrupt the normal mechanism at different stages). If we focus on the most common mutant found in those with cystic fibrosis in the USA, the Delta F 508 mutation, then the puzzling phenomenon is how does the CFTR protein misfold and get degraded (or not). Although some of the details of the degradation mechanism are still black boxes, nonetheless, we know that the outcome is that some misfolded proteins are found in the cells. Hence, enough of the mechanism is understood to direct empirical or rational drug discovery efforts, which may find a way to correct the misfolding and transport the protein to the cell membrane. The goal is to elicit sufficient amounts of chloride ion transport to restore some of the normal function and alleviate some of the disease symptoms.

However, if we want to know the mechanism by which this mutant leads to lung disease and death in CF patients, then there are still many black boxes to be filled. Competing hypotheses have to be evaluated about crucial stages of different mechanisms. To decrease death due to bacterial infections, it may be possible to direct therapeutic effects to the regulation of overexpression of neutrophils rather than correcting the CFTR gene itself. A different mechanism, coming later in the progression of the disease, becomes the target mechanism for controlling one disease symptom.

This case shows that the mechanistic perspective adds much more detail than a simple claim that a mutated gene causes the disease cystic fibrosis. That vague claim has been eliminated in favor of a rich description of the many mechanisms involved. One would have thought that for a disease due primarily to a single gene defect, we could say that the mutation in the gene causes the disease and the way to fix it is to apply gene therapy to deliver a functional, non-mutated gene. Sadly such a simple fix did not work. This case shows the importance of knowing the different stages of the normal mechanism and the specific ways in which it breaks and even identifying different mechanisms that come into play as the disease progresses. All these aid drug discovery.

5 Conclusion

In our "Thinking about Mechanisms" paper (MDC 2000), we discussed the simple relation between one puzzling phenomenon and one mechanism. One might have thought one could easily identify the phenomenon as the effect (E) and the entire mechanism (or some piece of it) that produces that phenomenon as the cause (C). However, there are many candidates for what is to be designated as the cause and what is to be called the effect, once specific features of a mechanism and its setup and start conditions are identified. The cystic fibrosis case illustrates advantages and complexities gained by discovering the relevant mechanisms, given the goals of explanation, prediction, and control over disease in medicine.

Acknowledgments I thank Hsiang-Ke Chao and the other organizers of the Taiwan Conference on the Philosophy of Biology and Economics at National Tsing Hua University in Hsinchu, March 24–25, 2011, for their invitation to present this paper and to all the participants for their comments on an earlier version.

My undergraduate research assistant, Sara Moghaddam-Taaheri, assisted my research on cystic fibrosis. I also thank Susan Lindee and Miriam Solomon for helpful information and useful references and Mitchell Drumm for scanning and sending his paper on the discovery of the CFTR gene. Robert Ennis, Tudor Baetu, Blaine Ford, Nancy Hall, Joan Straumanis, Eric Saidel, Justin Garston, Roberta Millstein, and an anonymous reviewer provided helpful comments on earlier drafts.

My research program on mechanisms in biology is aided by discussions with the Maryland Mechanisms Group and the DC History and Philosophy of Biology discussion group. My work on mechanisms profits from delightful collaboration with Peter Machamer and Carl Craver, whose ideas influenced this paper more than the citations indicate.

This research has been made possible in part by a grant from the US National Endowment for the Humanities: "Because democracy demands wisdom." Any views, findings, conclusions, or recommendations expressed in this paper do not necessarily represent those of the National Endowment for the Humanities. This research was also supported by sabbatical leave from the University of Maryland.

References

Anscombe, Gertrude Elizabeth Margaret. [1971] 1981. Causality and determination. In *Metaphysics and the philosophy of mind, the collected papers of G. E. M. Anscombe*, vol. 2, 133–147. Minneapolis: University of Minnesota Press.

Barros, D. Benjamin. 2008. Natural selection as a mechanism. *Philosophy of Science* 75: 306–322.

Bechtel, William. 2006. *Discovering cell mechanisms: The creation of modern cell biology*, Cambridge Studies in Philosophy and Biology. New York: Cambridge University Press.

Bechtel, William and Adele Abrahamsen. 2005. Explanation: A mechanist alternative. In ed. Carl F. Craver and Lindley Darden, Special Issue: "Mechanisms in Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton: Princeton University Press.

Bechtel, William, and Robert C. Richardson. 2010. *Discovering complexity: Decomposition and localization as strategies in scientific research*, 2nd ed. Cambridge, MA: MIT Press.

Bogen, James. 2004. Analysing causality: The opposite of counterfactual is factual. *International Studies in the Philosophy of Science* 18: 3–26.

Bogen, James. 2005. Regularities and causality; generalizations and causal explanations. In ed. Carl F. Craver and Lindley Darden, Special Issue: "Mechanisms in Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 397–420.

Bogen, James. 2008. Causally productive activities. *Studies in History and Philosophy of Science* 39: 112–123.

Bridges, Robert J. 2003. Pharmacology of delta F 508-CFTR biosynthesis. In *The cystic fibrosis transmembrane conductance regulator*, ed. Kevin L. Kirk and David C. Dawson, 181–200. New York: Kluwer.

Chmiel, James F., and Pamela B. Davis. 2003. Inflammatory responses in the cystic fibrosis lung. In *The cystic fibrosis transmembrane conductance regulator*, ed. Kevin L. Kirk and David C. Dawson, 160–180. New York: Kluwer.

Craver, Carl F. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Craver, Carl F. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Craver, Carl F., and William Bechtel. 2007. Top-down causation without top-down causes. *Biology and Philosophy* 22: 547–563.

Craver, Carl F., and Lindley Darden. 2001. Discovering mechanisms in neurobiology: The case of spatial memory. In *Theory and method in the neurosciences*, ed. Peter Machamer, R. Grush, and P. McLaughlin, 112–137. Pittsburgh: University of Pittsburgh Press. Reprinted in Darden (2006, Ch. 2).

Curlee, Kimberly V., and Eric J. Sorscher. 2003. Gene therapy for cystic fibrosis. In *The cystic fibrosis transmembrane conductance regulator*, ed. Kevin L. Kirk and David C. Dawson, 201–211. New York: Kluwer.

Darden, Lindley. 1991. *Theory change in science: Strategies from Mendelian genetics*. New York: Oxford University Press.

Darden, Lindley. 2001. Discovering mechanisms: A computational philosophy of science perspective. In *Discovery science*, Proceedings of the 4th International Conference, DS2001, ed. Klaus P. Jantke and Ayumi Shinohara, 3–15. New York: Springer.

Darden, Lindley. 2005. Relations among fields: Mendelian, cytological and molecular mechanisms. In ed. Carl F. Craver and Lindley Darden, Special Issue: "Mechanisms in Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 349–371. Reprinted in Darden (2006, Ch. 4).

Darden, Lindley. 2006. *Reasoning in biological discoveries: Mechanisms, interfield relations, and anomaly resolution*. New York: Cambridge University Press.

Darden, Lindley, and Carl F. Craver. 2002. Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences* 33: 1–28. Corrected and reprinted in Darden (2006, Ch. 3).

Drumm, Mitchell L. 2001. The race to find the cystic fibrosis gene: A Trainee's inside view. In *Cystic fibrosis in the 20th century: People, events, and progress*, ed. Carl F. Doershuk, 79–92. Cleveland: AM Publishing.

Eisenhaber, Frank (ed.). 2006. *Discovering biomolecular mechanisms with computational biology*. New York: Springer.

Glennan, Stuart S. 1996. Mechanisms and the nature of causation. *Erkenntnis* 44: 49–71.

Glennan, Stuart S. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69 (Proceedings): S342–S353.

Glennan, Stuart S. 2010. Ephemeral mechanisms and historical explanation. *Erkenntnis* 72: 251–266. doi:10.1007/s10670-009-9203-9.

Gu, Yuan Yuan, et al. 2009. Identification of *IFRD1* as a modifier gene for cystic fibrosis lung disease. *Nature* 458: 1039–1042.

Hedström, Peter. 2005. *Dissecting the social: On the principles of analytical sociology*. Cambridge: Cambridge University Press.

Kerem, Bat-sheva, et al. 1989. Identification of the cystic fibrosis gene: Genetic analysis. *Science* 245: 1073–1080.

Kirk, Kevin L., and David C. Dawson (eds.). 2003. *The cystic fibrosis transmembrane conductance regulator*. New York: Kluwer.

Knol, K. 1995. Cystic fibrosis: The past 25 years. *The Netherlands Journal of Medicine* 46: 266–270.

Konstan, Michael W., P.J. Byard, C.L. Hoppel, and P.B. Davis. 1995. Effect of high-dose ibuprofen in patients with cystic fibrosis. *The New England Journal of Medicine* 332(13): 848–854.

Lindee, Susan, and Rebecca Mueller. 2011. Is Cystic Fibrosis Genetic Medicine's Canary? *Perspectives in Biology and Medicine* 54(3): 316–331.

Liu, X., et al. 2002. Partial correction of endogenous delta F508 CFTR in human cystic fibrosis airway epithelia by spliceosome-mediated RNA trans-splicing. *Nature Biotechnology* 20(1): 47–52.

Liu, X., et al. 2005. Spliceosome-mediated RNA trans-splicing with recombinant adeno-associated virus partially restores cystic fibrosis transmembrane conductance regulator function to polarized human cystic fibrosis airway epithelial cells. *Human Gene Therapy* 16(9): 1116–1123.

Machamer, Peter. 2004. Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science* 18: 27–39.

Machamer, Peter, Lindley Darden, and Carl F. Craver [MDC]. 2000. Thinking about mechanisms. *Philosophy of Science* 67: 1–25. Reprinted in Darden (2006, Ch. 1).

Moghaddam-Taaheri, Sara. 2011. Understanding pathology in the context of physiological mechanisms: The practicality of a broken-normal view. *Biology and Philosophy* 26: 603–611.

Morgan, Thomas H., A.H. Sturtevant, H.J. Muller, and C.B. Bridges. 1915. *The mechanism of Mendelian heredity*. New York: Henry Holt & Company.

Pearson, Helen. 2009. Human genetics: One gene, twenty years. *Nature* 460: 164–169.

Quinton, Paul. 1983. Chloride impermeability in cystic fibrosis. *Nature* 301: 421–422.

Riordan, John R., et al. 1989. Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* 245(4922): 1066–1073.

Rommens, Johanna M., et al. 1989. Identification of the cystic fibrosis gene: Chromosome walking and jumping. *Science* 245(4922): 1059–1065.

Rowe, Steven M., Stacey Miller, and Eric J. Sorscher. 2005. Cystic fibrosis: Review article on mechanisms of disease. *The New England Journal of Medicine* 352(19): 1992–2001.

Skipper, Robert A. Jr., and Roberta L. Millstein. 2005. Thinking about evolutionary mechanisms: Natural selection. In ed. Carl F. Craver and Lindley Darden, Special Issue: "Mechanisms in Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 327–347.

Thagard, Paul. 1998. Explaining disease: Causes, correlations, and mechanisms. *Minds and Machines* 8: 61–78.

Thomson, Jeremy. 2002. Gene therapy hope for cystic fibrosis: SMaRT treatment for fatal disease shows promise. *Nature News*. doi:10.1038/news020101-6. Published online 4 January 2002. http://www.nature.com/news/2002/020104/full/news020101-6.html. Accessed 22 Sept 2011.

Wang, X., et al. 2006. Hsp90 cochaperone aha1 downregulation rescues misfolding of CFTR in cystic fibrosis. *Cell* 127: 803–815.

Watanabe, Akira, Seungho Choe, Vincent Chaptal, John M. Rosenberg, Ernest M. Wright, Michael Grabe, and Jeff Abramson. 2010. The mechanism of sodium and substrate release from the binding pocket of vSGLT. *Nature* 468: 988–992.

Watson, James D., Tania A. Baker, Stephen P. Bell, and Alexander Gann. 2007. *Molecular biology of the gene*, 6th ed. San Francisco: Benjamin Cummings.

Widdicombe, J.H. 2003. CFTR and airway pathophysiology. In *The cystic fibrosis transmembrane conductance regulator*, ed. Kevin L. Kirk and David C. Dawson, 137–159. New York: Kluwer.

Woodward, James. 2002. What is a mechanism? A counterfactual account. *Philosophy of Science* 69(4): S366–S377.

Chapter 3
Identity, Structure, and Causal Representation in Scientific Models

Kevin D. Hoover

Abstract Recent debates over the nature of causation, casual inference, and the uses of causal models in counterfactual analysis, involving *inter alia* Nancy Cartwright (*Hunting Causes and Using Them*), James Woodward (*Making Things Happen*), and Judea Pearl (*Causation*), hinge on how causality is represented in models. Economists' indigenous approach to causal representation goes back to the work of Herbert Simon with the Cowles Commission in the early 1950s. The paper explicates a scheme for the representation of causal structure, inspired by Simon, and shows how this representation sheds light on some important debates in the philosophy of causation. This structural account is compared to Woodward's manipulability account. It is used to evaluate the recent debates – particularly, with respect to the nature of causal structure, the identity of causes, causal independence, and modularity. Special attention is given to modeling issues that arise in empirical economics.

1 Models and Causes

Formal scientific models possess some distinct advantages over verbal accounts. (There are, to be sure, disadvantages as well.) All representations (formal or verbal) are partial: they omit, simplify, approximate, and idealize; they fall short of saying

Prepared for the *International Conference on the Philosophy of Economics and Biology* at the National Tsing Hua University, Hsinchu, Taiwan, 24–25 March 2011; it is a substantial revision of a paper first presented at the conference on *Modeling the World: Perspectives from Biology and Economics*, Helsinki, 28–30 May 2009. I am grateful to François Claveau and two anonymous referees for comments on an earlier draft. I acknowledge the support of the US National Science Foundation (grant no. NSF SES-1026983).

K.D. Hoover (✉)
Departments of Economics and Philosophy, Duke University,
213 Social Sciences Building, Box 90097, Durham, NC 27708-0097, USA
e-mail: kd.hoover@duke.edu

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_3,

everything that could be truly said and short of saying everything that we might like to say. Recognition of the gap between the representation and the world leads various philosophers – among them, Paul Teller (2001) and Ronald Giere (2006) – to reconceptualize scientific knowledge as perspectival. Part of the reconceptualization is a repudiation of the view that *omni*science sets the standard for the worthiness of scientific knowledge.

The truth in a once-common vision of science is that the only fully adequate scientific knowledge trades in exceptionless, universal generalizations – scientific laws. All more specific knowledge is, in principle at least, derivable from these laws. Recognizing – as indeed any serious philosopher or scientist must – that we do not, in fact, possess all the laws simply meant that what we did possess was a slightly shabby, deficient version of what we wanted. We do not stand on Olympus, but science was nonetheless to be judged from the Olympian heights.

An alternative vision of science championed by Giere and Teller, as well as by Nancy Cartwright (1999), and William Wimsatt (2007), among others, starts lower and builds upward. The standards of good or successful science are partial and local, and science itself is constructed, to use an apt term from the subtitle of Wimsatt's (2007) book, in a *piecewise* manner.

The local knowledge that grounds science in this vision is often causal knowledge. Yet, like other parts of science, causation has often been analyzed top down. Many accounts of causation – for example, those of David Lewis (1973) and Daniel Hausman (1998) – explicate causes against a background of universal laws. In contrast, piecewise accounts of science typically take the causal relation as primitive or, at least, built from something more local and specific than universal laws.

A piecewise approach is especially suited to economics and other social sciences, biology, and areas of physical sciences, such as climatology – fields that would be hard to analyze from a small set of universal laws on the model of Newtonian mechanics. Economists, for example, increasingly conceptualize economics causally, as evident in the work of Clive Granger in time-series econometrics, James Heckman in microeconometrics, recent developments in "natural experiments" in economics, and counterfactual analysis.[1]

Causal realism is the doctrine that causal relationships exist in the world and that the role of causal models is to represent them adequately for some purpose. Not all scientists (nor all philosophers) who talk about causes are realists, but that is a question for another day. Here, I want to focus on *representation* of causes and not on fundamental ontology or epistemology. It is a commonplace that different representations or notational schemes allow us to see different things and that some schemes are more effective than others – consider Arabic numerals or Feynman diagrams. The main goal of this chapter is to develop a scheme for representing causal relationships and to consider the light that it sheds on how we

[1] See Hoover (2008 and 2012a) on the place of causal analysis in economics and Reiss (2007) on natural experiments and counterfactual analysis. Hoover (2004) documents the fall and rebirth of causal analysis and language in economics.

should understand causation generally. The roots of the approach advocated here are found in my own work as a practitioner of economics and draw on sources, such as the work of Herbert Simon, that were originally aimed at problems that arose in economic and econometric analysis. The application is much broader than these origins might suggest.

In part, this chapter reacts to Cartwright's (2007) "pluralistic" account of causation. In stressing plurality, Cartwright fails to illuminate the close relationships among a number of approaches to causality that are hidden in alternative schemes of representing causal relations. In part, the chapter reacts to James Woodward's (2003) "manipulability" account of causation – an account which is much criticized by Cartwright. Woodward's understanding of causation appears to be driven by particular schemes of representing causes. A more effective scheme of representation suggests different conclusions with respect to several important issues. The account proposed here in no way fundamentally conflicts with the general approach of modeling causal relationships graphically, developed especially by Judea Pearl (2000) and Peter Spirtes et al. (2000) and used by Woodward. Rather it clarifies the relationship between graphical representations and systems of equations in a manner that both enriches the graphical approach and demonstrates the fundamental kinship of the two approaches.

2 Representing Causal Structure

2.1 Graphs and Equations

While many philosophers understand causal relations as holding fundamentally among particular events, occurrences, or properties (i.e., among *tokens*), Woodward and most economists understand causal relations as holding among variables (i.e., among *types*). Token-level relationships for Woodward and the economists are causal to the degree that they instantiate a type-level relationship. In stochastic cases, token-level relationships are seen as the realization of random processes. Relations among variables are often expressible in the form of systems of equations. Equality is a symmetrical relationship, and the most distinctive characteristic of causal relations is their asymmetry: A causes B gives no ground for holding that B causes A (although we must not rule out mutual causation without further consideration). Woodward, in common with Pearl (2000), Spirtes et al. (2000), and other advocates of graph-theoretic or Bayes net methods of causal inference, represents causal relations by graphs in conjunction with equations.

Figure 3.1 shows a typical causal graph (uppercase letters represent variables) that corresponds to a system of equations:

$$A = \alpha_A, \tag{3.1}$$

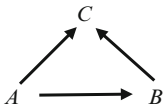


Fig. 3.1 Causal graph of the system (3.1)–(3.3)

$$B = \alpha_{BA}A, \tag{3.2}$$

$$C = \alpha_{CA}A + \alpha_{CB}B, \tag{3.3}$$

where, for the moment, we regard the α_{ij} as fixed coefficients.

Systems of equations are causally ambiguous. In stochastic cases, we generally recognize that correlation is not causation; in nonstochastic cases, the analogue is that functional relations are not causation. The arrows in the graph represent the primitive notion of the asymmetry of causation.

Graphs and equations interpreted causally both have a long history in economics (see Hoover 2004). But it is fair to say that equations have gained the upper hand and that, for many years, causation itself was rarely referred to directly, but at best was implicit in distinctions between dependent and independent (or endogenous and exogenous) variables and in synonyms and circumlocutions: instead of "*A* causes *B*," *A produces*, *influences*, *engenders*, *affects*, or *brings about B*, or *B reflects*, *is a consequence of, is a result of*, or *is an effect of A* (see Hoover 2009). A growing wariness of causal language went hand in hand with a wariness of graphical representation. As Pearl puts it:

> Early econometricians were very careful mathematicians; they fought hard to keep their algebra clean and formal, and they could not agree to have it contaminated by gimmicks such as diagrams. (Pearl 2000, p. 347)

Equations alone are causally ambiguous, since in themselves they do not represent causal asymmetry. But graphs are themselves causally ambiguous, because quite different functional relationships can be represented by the same graph (Woodward 2003, p. 44). Just as economists found circumlocutions to express "cause," they have typically – although not necessarily consistently – represented causal asymmetry by the convention of writing causes on the right-hand side and effects on the left-hand side of equations. Various devices have been suggested for explicitly combining the functional detail of systems of equations with the asymmetries of the graph. In lieu of the equal sign, Cartwright (2007, p. 13) suggests a causal equality ($c^{=}$) which Hoover (2001, p. 40) writes as ($\Leftarrow$). With a new notational device, the graph in Fig. 3.1 could be omitted and the system (3.1), (3.2), and (3.3) could, then, be rewritten as

$$A \Leftarrow \alpha_A, \tag{3.1'}$$

$$B \Leftarrow \alpha_{BA}A, \tag{3.2'}$$

$$C \Leftarrow \alpha_{CA}A + \alpha_{CB}B. \qquad (3.3')$$

Cartwright (2007, p. 16, *passim*) refers to such equations as "causal laws" – that is, laws that connect specific causes to specific effects. Woodward (2003), as well as most of the literature on graphical causal models, considers one-way causation only. Economists refer to such systems as *recursive*, while the graphical representations are often known as *directed acyclical graphs* (*DAG*s). Cyclical graphs (e.g., $A \rightarrow B \rightarrow C \rightarrow A$) are sometimes entertained, but the tight cycle of the simultaneous system ($A \rightarrow B \rightarrow A$ or $A \leftrightarrow B$), a bread-and-butter system in economics, is encountered far less frequently. The equations themselves are generally taken to be linear – especially linear in parameters. While these restrictions are by no means necessary, they highlight the inadequacy of the graphs fully to represent various levels of causal complexity. Economists avoiding graphs (*pace* Pearl) are perhaps partly motivated by an appreciation of the subtlety of causal representation and not some intuitive revulsion toward graphical gimmickry.

2.2 Simon on Causal Order

Following Haavelmo's "The Probability Approach in Econometrics" (1944), econometricians focused on what Frisch had called the "inversion problem" – namely, how to infer the original structure from passive observation of the data that it generates (Louçã 2007, p. 95). Later dubbed the "identification problem," a detailed account of the mathematics was for a time the central focus of the Cowles Commission (Koopmans 1950; Hood and Koopmans 1953). Identification naturally requires something to identify. Simon's contribution to the 1953 Cowles Commission volume sought to characterize the causal order of a system of equations.

Simon started with a *complete* system of equations – that is, a system that could be represented as a multivariate function with a well-defined solution. He then focused on *self-contained subsystems* of the complete system. To illustrate, Eqs. (3.1), (3.2), and (3.3) form a complete system. Equation (3.1) is a self-contained subsystem in that it determines the value of A without reference to any other equation. Equations (3.2) and (3.3) considered separately are not self-contained subsystems as they do not contain enough information to determine B or C. In contrast, Eqs. (3.1) and (3.2) together are a self-contained subsystem, since they determine the values of A and B without reference to Eq. (3.3).

Simon's conception is closely related to his later work on hierarchies of systems (Simon 1996; see also Hoover 2012c). Causes are the outputs of lower-level systems and the inputs to higher-level systems. The relationship is closely connected to the solution algorithms for systems of equations. In system (3.1), (3.2), and (3.3), A is determined entirely by (3.1) and can be regarded as an output. If we know A, we do not need to know (3.1) to determine B; a specific value for A forms an input that, in effect, turns the non-self-contained subsystem (3.2) into a

self-contained subsystem. Its output is, of course, B. Knowing B alone, however, does not turn (3.3) into a self-contained subsystem. Substituting its value into (3.3) leaves the variable A in place (despite the fact that B cannot have a well-defined value unless A also has a well-defined value), and we have to substitute A directly from (3.1). Thus, A directly causes B, and A and B directly cause C; so, A is both a direct and an indirect cause of C. This, of course, is the causal structure of Fig. 3.1.

Suppose, however, that we modify the system slightly by replacing Eq. (3.3) with (3.3″):

$$C = \alpha_{CB} B. \tag{3.3''}$$

Then, A directly causes B, and B directly causes C, but A only *indirectly* causes C: Eq. (3.1) is nested in the complete system (3.1), (3.2), (3.3″), but the self-contained subsystem (3.1) and (3.2) intervenes between the self-contained subsystem (3.1) and the self-contained complete system (3.1), (3.2), (3.3″).

Simon's analysis assumes that the original way of writing the equations is canonical. But he notices that the same functional relationships can be represented by other sets of equations. So, for example, the self-contained subsystem (3.1) and (3.2) could be replaced by

$$A = \beta_A + \beta_{AB} B, \tag{3.4}$$

$$B = \beta_B, \tag{3.5}$$

which has the same numerical solution as (3.1) and (3.2) provided that

$$\beta_A = \frac{\alpha_A}{(1 - \alpha_{BA})}, \tag{3.6}$$

$$\beta_{AB} = \frac{-1}{(1 - \alpha_{BA})}, \tag{3.7}$$

and

$$\beta_B = \alpha_A \alpha_{BA}. \tag{3.8}$$

(Nothing depends on the fact that the β_{ij} are defined in terms of the α_{ij}. We could as easily have started with Eqs. (3.4) and (3.5) and derived an analogous set of restrictions defining the α_{ij} in terms of the β_{ij} to guarantee identical solutions.) The two sets of equations have the same solution, but under Simon's analysis B causes A in (3.4) and (3.5), whereas A causes B in (3.1) and (3.2). Indeed, since every linear combination of Eqs. (3.1) and (3.2) is functionally equivalent, we can easily write down systems that would be interpreted as having no causal connections or as displaying mutual causation. This is the sense in which systems

of equations are causally ambiguous, which is the rationale for supplementing them with graphs.

Simon does not appeal to graphs. Instead, he considers a higher-order relation of *direct control* over parameters (Simon 1953, pp. 24–27). He invites us (and nature) to experiment on a system by directly controlling the value of its parameters (the coefficients now being thought of as parameters that can take different values). The privileged parameterization is the one in which such experiments can be conducted independently. Thus, if one represents a causal system by Eqs. (3.1) and (3.2) and can control A directly by choosing α_A and thereby control B indirectly without altering the functional form of Eq. (3.2), then the parameter set $\{\alpha_A, \alpha_{BA}\}$ is privileged. No other functionally equivalent system shares this property.

If, for example, (3.1) and (3.2) represented the true causal order, but we instead modeled the causal relationships with (3.4) and (3.5), our control of A and B would not show the same sort of functional invariance. In fact, the only way to achieve the same values for A and B would be for the coefficient values of $\{\beta_A, \beta_B, \beta_{AB}\}$ to shift according to the restrictions (3.6), (3.7), and (3.8). In effect, the decision that $\{\alpha_A, \alpha_{BA}\}$ is *the* parameter set – and that any other set of coefficients (e.g., $\{\beta_A, \beta_B, \beta_{AB}\}$) are simply functions of those parameters – determines the causal direction among the variables: it puts the arrowheads on the shafts.

2.3 The Structural Account of Causal Order

I refer to an account of causal order based on Simon's seminal analysis as the *structural account*.[2] It is structural in the sense that what matters for determining the causal order is the relationship among the parameters and the variables and among the variables themselves. The parameterization – that is, the identification of privileged set of parameters that govern the functional relationships – is the source of the causal asymmetries that define the causal order. The idea of a privilege parameterization can be made more precise, by noting that a set of parameters is privileged when its members are, in the terminology of the econometricians, variation-free. A parameter is *variation-free* if, and only if, the fact that other parameters take some particular values in their ranges does not restrict the range of admissible values for that parameter.

Defining parameters as variation-free variables has a similar flavor to Hans Reichenbach's (1956) *Principle of the Common Cause*: any genuine correlation among variables has a causal explanation – either one causes the other, they are mutual causes, or they have a common cause. Since we represent causal connections as obtaining only between variables *simpliciter*, we insist that parameters not display any mutual constraints. Whereas, the Principle of the Common Cause is a metaphysical or methodological presupposition with significant bite, the variation-freeness of

[2] A more formal presentation of the structural account is given in Hoover (2001, Chap. 3).

parameters is only a representational convention. Any situation in which it appears that putative parameters are mutually constraining can always be rewritten so that the constraints are moved into the functional forms that connect variables to each other.

For example, in the system

$$X = a \tag{3.9}$$

$$Y = bX, \quad b \leq a \tag{3.10}$$

the parameters are not variation-free, since the choice of a constrains the value of b. However, this system can be reformulated into a related (nonlinear) system with the same solutions in which the parameters are variation-free:

$$X = a \tag{3.11}$$

$$Y = \begin{cases} bX, & \text{if } a \geq b \\ \textit{undefined}, & \text{if } a < b. \end{cases} \tag{3.12}$$

Because of its analogy with the Principle of the Common Cause, we refer to the stipulation that parameters be variation-free as the *Reichenbach Convention*.

Except for the system of Eqs. (3.11) and (3.12), we have considered only linear equations. But the structural account can accommodate nonlinearity quite generally. The key step is that parameters are not defined as coefficients uniquely associated with particular variables, as they are, for example, in path analysis, in which the parameters are merely the regression weights associated with each causal arrow.

To see the role of nonlinearity, consider a simplified example of a two-equation system from a macroeconomic model with rational expectations[3]:

$$m_t = \lambda + m_{t-1} + \varepsilon_t, \tag{3.13}$$

$$p_t = m_t + \alpha\lambda - \delta + \nu_t. \tag{3.14}$$

The subscripts are time indices. Our concern is only with the causal relationship between m_t and p_t, so the lagged value of m can be regarded as a constant. While it is not vital for our purposes, (3.13) is interpreted as a rule for fixing the money supply, while (3.14) determines the price level.

It is obvious that in Simon's framework m_t directly causes p_t. In our earlier examples, there was a simple, natural association of individual parameters with individual variables in equations written in a canonical form (causes on the right-hand side; the effect on the left-hand side; the two sides connected by an

[3] The model is drawn from Hamilton (1995).

asymmetrical assignment operator ("⇐")).[4] But here we cannot associate the parameter λ exclusively with either equation. Indeed, the notion of a canonical form of equations is merely heuristic and must be abandoned in this case.

Equations displaying this sort of nonlinearity in parameters are referred to in the macroeconometrics literature as subject to "cross-equation restrictions." Suppose that the monetary authority wants to loosen monetary policy; it would increase λ. Because of the cross-equation restriction, in addition to the direct causal effect of m_t on p_t, there is a change in the functional relationship between p_t and m_t. In a stochastic version of the model, the conditional probability distribution of p_t on m_t would not be invariant to changes in m_t. This striking conclusion is well known to economists as the "Lucas critique" (Lucas 1976).[5] Economists often discuss it in terms of "deep parameters" (here α and λ) versus empirically observable coefficients (say, a regression coefficient Π, which in fact equals $\alpha\lambda - \delta$, but which is estimated as a unit). In terms of our account of causal representation, the deep parameters are just the parameters that define causal order.

While the Lucas critique is not unknown to philosophers, it is not always appreciated that it undermines any necessary connection between a well-defined causal relationship and the invariance of the probability of an effect *conditional* on its causes. Indeed, our account of causal order suggests that it is the invariance of the probability distribution of the cause (the *marginal* probability distribution) to independent changes of other causes of the effect that is the empirical hallmark of a causal relation (see Hoover 2001, Chap. 8). This claim amounts to saying that it is not the conservation of the functional relationship of causes to effect as causes vary that is most characteristic of causal relations; rather it is that effects do not flow backward against the causal arrow.

2.4 Causal Identity

Implicit in our discussion so far is the notion that variables in causal relationships must be causally distinct. Let us make this notion more explicit. Variables are distinguishable when we have some independent means of measuring, observing, or characterizing them. Yet variables that are distinguishable in this general way need not be causally distinct.

To take an economic example, prices (P) are distinguishable from quantities (Q), but consider the simple supply and demand model in which quantities and prices are mutually determined:

[4] Which in fact suggested the scheme of distinguishing parameters by subscripts: for example, α_{BC} was the parameter multiplying the variable C in the canonical equation for B.

[5] For expositions of the Lucas critique, see Hoover (1988, Chap. 8, section 8.3; 2001, Chap. 7, section 7.4).

$$Q = \alpha + \beta P \tag{3.15}$$

$$P = \delta + \gamma Q. \tag{3.16}$$

Solving (3.15) and (3.16) yields

$$Q = \frac{\alpha + \beta\delta}{1 - \beta\gamma}, \tag{3.17}$$

$$P = \frac{\delta + \alpha\gamma}{1 - \beta\gamma}. \tag{3.18}$$

Both variables are determined by the same set of parameters. It would be impossible, therefore, that we alter the value of one of them without also altering the value of the other. We might, then, regard the two variables as having a two-way or mutual causal relationship. But should we really call variables that have no causal relationships distinguishable from one another as standing in a causal relationship with each other? It would be more to the point to say that, causally speaking, there is no difference between them.

The issue arises not only in simultaneous systems of equations. Consider instead the following system:

$$A = \alpha, \tag{3.19}$$

$$B = \beta A, \tag{3.20}$$

$$C = \beta A. \tag{3.21}$$

On Simon's criterion, A clearly causes both B and C, but what is the causal relationship between B and C? It might appear to be mutual, since there is no intervention on either variable that does not alter the other. But this seems counter-intuitive, because the connection is through the parameter β rather than through the variables; yet our presumption is that causal relationships are mediated only through variables. If we imagine the variable B and Eq. (3.20) eliminated, nothing would change for C.

We see, then, that some systems with or without mutual or simultaneous causation are problematic, but there is no reason to believe that problematic cases arise inevitably in simultaneous systems. We need a way of characterizing problematic and unproblematic systems. This suggests that we characterize causal identity and causal distinctiveness:

Causal Identity: Two variables are *causally identical* if aside from their mutual relationship, they have all the same causes and effects.

Causal Distinctiveness: Variables that are not causally identical are *causally distinct*.

In invoking causes and effects, these definitions are not circular, since whether or not the relevant causal relationships exist can be determined from the parameterization of the system in line with our elaboration of Simon's structural account.[6]

When variables are distinguishable because we possess independent means of measuring, observing, or characterizing them, a failure also to be causally distinct will be rare. Causal identity is more likely to be a property of an impoverished representation of the world, arising most naturally in cases in which a few variables stand in a tight relationship. Causally, identity will rarely arise in nondeterministic cases, as the variables that describe such cases are, in general, subject to "shocks" that distinguish one from another. However, models are nearly always highly simplified, and shocks that are small enough in the world may be neglected in a model, so that, if the world produces "near causal identity," a good model of the world may produce exact causal identity (cf. Suppes 1970, p. 33 on ϵ-direct cause). Similarly, in selecting a simplified representation of the world, we may choose to ignore some ways in which variables could be causally distinguished – again, producing causal identity in the model.

There is one type of case of causal identity that is not rare, but a pitfall to be carefully avoided. Conceptual identities or variables that are "equal by definition" belong to a special class of causal identity that does not depend on modeling choices but on the meaning of the variables. For example, the price (per dollar of coupon payment) of a perpetual bond or consol (P_C) is, by definition, the inverse of its yield (R): $P_C \equiv 1/R$. Anything that affects the yield affects the price; yet we should not regard these variables – conceptually different and with different units of measurement – as causally related. A system of equations that embedded this identity would, according to our definitions, find P_C and R to be causally identical and, therefore, not stand in any other sort of causal relationship with each other. This is exactly as it should be.

3 The Structural Account Versus the Manipulability Account of Causation

So far the discussion of causal order has been formal. But the importance of a representational scheme arises from its power to illuminate genuine scientific issues, a matter to which we now turn.

[6] Causal identity can be thought of as a metaphysical property of the world and as a property of a model or representation. Elsewhere I have argued in favor of a *perspectival realism* in which a successful model tells us the truth about the world from a particular point of view (Hoover 2012b, see also 2012c), which reduces the force of a distinction between the metaphysics and the properties of the model.

3.1 Modularity

I have referred to the development of Simon's scheme of causal representation as the *structural account* of causation. While I will not discuss it in detail here, a causal structure is, I believe, essentially what some philosophers mean when they refer to *mechanisms*. A causal model is thus the representation of the workings of a mechanism. The account is "structural," in a formal sense, in that causal order depends on nonunique functional relationships among variables acquiring a unique form or structure through the specification of the parameter space. But what are the appropriate semantics? That is, how in reference to the world – that is, in reference not to the representation but to what it represents – should we understand the parameters? A natural reading, suggested by Simon's notion of an experimenter's ability to control or intervene directly to set parameter values, is to regard parameters as the loci of interventions. Such an interpretation points to a similarity to Woodward's (2003) *manipulability account* of causation. While the similarity is genuine, we should distinguish the structural account from the manipulability account.

Although Woodward (2003, Chap. 2) provides a detailed and nuanced development of the manipulability account, the essential point is conveyed in his definition of a *direct cause*:

> (**DC**) A necessary and sufficient condition for X to be a direct causes of Y with respect to some variable set V is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in V besides X and Y are held fixed at some value by interventions. (Woodward 2003, p. 55)[7]

Despite Woodward (2003, p. 39) regarding causation as fundamentally a type-level relationship among variables, (**DC**) defines direct cause in terms of a token-level action – an intervention. For example, an intervention on the variable B in Fig. 3.1 would set B to a particular value, say, $B = b$, and holding it fixed at that value amounts to wiping out or breaking the arrow from A to B, indicating that no change in A is allowed to affect B. B is a direct cause of C, according to (**DC**) if C changes (or would change, the intervention being conceived of counterfactually) as a result of this intervention.

Pearl (2000, p. 70) represents interventions by the operators "set(X)" or "do(X)." Woodward (2003, pp. 47–48) notes "X and set X are not really different variables, but rather the same variable embedded in different causal structures. . ." After the intervention, we can represent Fig. 3.1 with a new graph as in Fig. 3.2. The transition from one graph to the other – from one causal structure to another – presupposes that the wiping out of causal arrows without affecting other parts of the graph makes sense. Woodward refers to the property that warrants such an intervention as *modularity*:

[7] I have written V where Woodward writes **V**, to remain consistent with the notation of Sect. 2 above.

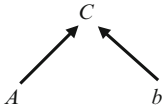


Fig. 3.2 Causal graph of Fig. 3.1 after the intervention set ($B = b$)

> a system of equations will be modular if it is possible to disrupt or replace (the relationships represented by) any one of the equations in the system by means of an intervention on (the magnitude corresponding to) the dependent variable in that equation, without disrupting any of the other equations. (Woodward 2003, p. 48)

And while he recognizes that representations of causal relationships may not always display modularity, he assumes

> that when causal relationships are correctly and fully represented by systems of equations, each equation will correspond to a distinct causal mechanism and that the equation system will be modular. (Woodward 2003, p. 49)

Cartwright (2007, part II) objects to modularity as an essential feature of causation.[8] The structural account illuminates both what is right and what is wrong in Cartwright's objections. Cartwright denies that all well-defined causal systems are modular. And she is correct. We should notice, first, that the system defined by Eqs. (3.13) and (3.14), which has a well-defined causal order on the structural account, is itself not modular, since the individual equations do not represent distinct mechanism, but can function only as a pair (see Hoover, 2011, section 16.3.2). Cartwright herself argues largely through counterexamples. The first example is a carburetor (Cartwright 2007, pp. 15–16). Cartwright describes the operation of the carburetor through a system of equations in which key coefficients depend on the geometry of its chamber[9]:

> we can see a large number of [functional] laws all of which depend on the same physical features – the geometry of the carburettor. So no one of these laws can be changed on its own. To change any one requires a redesign of the carburettor, which will change the others in train. By design the different causal laws are harnessed together and cannot be changed singly. So modularity fails. [Cartwright 2007, p. 16]

Cartwright illustrates her point with a set of causal laws from which I reproduce two (in an altered notation):

$$X \Leftarrow h(G, A; \gamma), \tag{3.22}$$

$$A \Leftarrow j(S, T; \sigma), \tag{3.23}$$

where $X =$ gas exiting the emulsion tube; $G =$ gas in the emulsion tube; $A =$ air pressure in the chamber; $S =$ suck of the pistons; $T =$ throttle valve; and

[8] Cartwright's chapters are her side of a vigorous debate over modularity carried on with Hausman and Woodward (1999, 2004).

[9] I write "coefficients," not "parameters" as Cartwright does, since she assumes that they are functions of other things, violating the usage established in Sect. 2.3 above.

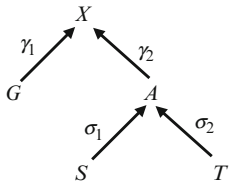


Fig. 3.3 The causal structure of Cartwright's carburetor

$$\gamma = \gamma(\text{geometry of chamber}, \ldots), \tag{3.24}$$

$$\sigma = \sigma(\text{geometry of chamber}, \ldots). \tag{3.25}$$

The system of equations can be represented in a causal graph as in Fig. 3.3. An oddity of Cartwright's exposition is that each of the principal equations has only one coefficient. I do not think, however, that any of her critical points hang on that, so I will treat them as vectors: $\gamma = [\gamma_1\ \gamma_2]$ and $\sigma = [\sigma_1\ \sigma_2]$. In Fig. 3.3, the elements of these vectors are listed alongside the appropriate causal arrows as indices of the strength of the influence of each cause over its effect.[10]

Cartwright's point is that modularity requires that each cause can be intervened upon separately. So, for example, if we wish to change A, we have to change σ. A change in σ can be achieved through a change in the geometry of the chamber, but that necessarily changes γ as well; so the causal relationship of G and A to X is neither distinct nor invariant with respect to that of S and T to A. Modularity fails. (As indeed it does in the closely analogous monetary-policy example of Eqs. (3.13) and (3.14).)

The representational conventions of the structural account force us to take a stand on some of the details of Cartwright's example. First, γ and σ are not parameters as we have defined the term (see fn. 8). Their interdependence violates the Reichenbach Convention. We must decide, then, whether they are variables or simply coefficients (a shorthand way of grouping parameters that interact with a variable when writing a function). Second, the "geometry of the chamber" is unlikely to be characterized by a single variable or parameter. In practice, the most natural way of representing it would be as a set of interrelated variables governed by parameters that conform to the Reichenbach Convention. Imagine representing the geometry in a computer-automated design program. The designer can set various parameters independently to generate various shapes that constitute the "geometry of the chamber." Aspects of that geometry (which can be represented as causally salient variables) are what feed into Cartwright's "causal laws" through γ and σ.

Figure 3.3, or even a more elaborate diagram, is too coarse to represent the refinements to the causal structure of the carburetor needed adequately to flesh out

[10] This makes the inessential, but in this case harmless, assumption that the equations are linear in variables.

Cartwright's account of its working. The rough sketch of a structural causal representation of the carburetor supports Cartwright's view that in this case, as in many others, a causally unambiguous system need not be modular.

Where Cartwright goes astray is in her belief that the failure of a well-defined causal system to be modular in Woodward's sense threatens an interventionist account and not just Woodward's particular formulation of it (the manipulability account). The structural account is a type of interventionist account that relies on a different sort of modularity – that is, modularity at the level of parameters. By definition, parameters can change independently of each other. Cartwright might object to the assumption that parameters are necessarily independent (or variation-free). But as I previously argued, this is a matter of convention; representations can always be formulated with variation-free parameters, constraints having been moved into the functional relationship of variables. As a *conventional* restriction on causal representation, however, modularity of the parameters does not pack any punch: it does not tell us that a causal mechanism can be disassembled into parts that operate independently of each other. Indeed, modularity of that sort is not conventional but substantial and highly special. The modularity of parameters is, nonetheless, the sort of modularity that we need to *define* causal structure. The structural account allows us to see that Woodward's definition of direct cause is too strong; it rules out too many relationships that are clearly causal in an obvious and practical sense.

Woodward may object to Cartwright's implicit, and the structural account's explicit, characterization of an intervention. For Woodward, an intervention is setting a variable to a value come what may – a severing of its relations to its own causes. In the structural account, an intervention is a more delicate matter of influencing a variable in some particular way by changing one or more parameters in a context in which multiple parameters connected to the variable under some functional constraints are the rule.

The failure of modularity does not depend on which notion of intervention we employ. Wiping out a causal arrow (or equation) does not necessarily leave other causal arrows intact. Consider the monetary-policy system (3.13) and (3.14) referred to in our earlier discussion of the Lucas critique. The cross-equation restriction (i.e., the appearance of λ in both equations) arises because of the assumption that agents form expectations of the path of the money supply (m_t) based on knowledge of the policy rule. Woodward's type of intervention would amount to setting m_t to a definite value independent of its past value – essentially wiping out the causal arrow from m_{t-1} to m_t. Eliminating that causal arrow does not merely imply a change in the values of the parameters of (3.14), which would be a failure of invariance of the sort highlighted by the Lucas critique and implicit in Cartwright's carburetor example, it would in fact render the parameter λ meaningless as it would undercut any basis for forming a rational expectation of the path of m_t. In effect, the wiping out of the causal arrow from m_{t-1} to m_t does not merely alter the causal arrow from m_t to p_t; it smashes it. There are plenty of real-world examples of devices in which one part cannot be removed without breaking others, which nonetheless possess well-defined causal structure.

Cartwright objects to Woodward defining direct cause by interventions that set a variable to a value come what may precisely because, as we already observed, such interventions alter the causal system (e.g., moving from Fig. 3.1 to Fig. 3.2), even when the other causal arrows (and the equations to which they correspond) are left intact. We should be concerned with the normal workings of a causal system, and the workings of some other causal system are irrelevant to them (Cartwright 2007, p. 107; Cartwright and Jones 1991).

Another of Cartwright's counterexamples to modularity – the operation of "a well-made toaster" – helps to clarify the point:

> The expansion of the sensor due to the heat produces a contact between the trip plate and the sensor. This completes the circuit, allowing the solenoid to attract the catch, which releases the lever. The lever moves forward and pushes the toast rack open.
>
> I would say that the bolting of the lever causes the movement of the rack. It also causes a break in the circuit. Where then is the special cause that affects only the movement of the rack? Indeed, where is there space for it? The rack is bolted to the lever. The rack must move exactly as the lever dictates. So long as the toaster stays intact and operates as it is supposed to, the movement of the rack must be fixed by the movement of the lever to which it is bolted.
>
> Perhaps, though, we should take the movement of the lever to the rack as an additional cause of the movement of the rack? In my opinion we should not. To do so is to mix up causes that produce effects within the properly operating toaster with the facts responsible for the toaster operating in the way it does; that is, to confuse the causal laws at work with the reason those are the causal laws at work. (Cartwright 2007, pp. 85–86)

What, we may ask, is *proper* operation? To ask such a question requires that we can distinguish changes in its state that constitute its proper operation from changes that undermine the proper operation or destroy the mechanism. Surely, such a distinction is partly a matter of perspective and often driven by pragmatic considerations. It requires that we be able to decide when the mechanism has been so altered that it is effectively a new mechanism and when the mechanism is preserved – that is, we need identity conditions for a causal system (cf. Woodward 2003, pp. 108–109).

Cartwright uses the toaster example to argue that the relevant interventions operate only within a context of a preserved mechanism and that Woodward's come-what-may interventions generally break the mechanism. While she does not provide the necessary identity conditions, they are evident in the structural account. Two mechanisms are causally the same when they have the same parameterization (i.e., the same privileged set of variation-free parameters) and differ only in the particular values that the individual parameters take within their admissible ranges. In other words, two mechanisms are causally identical when they differ not in their parameters or variables but in the token instantiations of their parameters and the token consequences of those instantiations for the variables.

Another example illustrates both the pragmatic and the conceptual issue. In the movie *The African Queen*, Charlie Allnut (Humphrey Bogart's character) runs a steamboat. From time to time the pressure in the boiler of the steam engine gets dangerously high. He hits a particular valve with a hammer, which frees up the valve, and allows the steam to escape. Later in the film, as part of his general effort

to make himself and the boat more presentable to Rose Sayer (Katharine Hepburn's character), he cleans and lubricates the valve so that the pressure is released automatically without the use of the hammer.

In each case, with or without the sticky valve, the steam engine has a *typical* operation. Which is proper depends as much on Charlie's relationship with Rose as on any fact about the steam engine. Charlie cleaned up is a different man; the steam engine cleaned up is different engine. But there are important senses in which both are still the same. In the case of the engine, if not the man, we can represent the preserved mechanism as one in which there is a parameter (or parameters) that governs whether the engine is in its clean or dirty state.

We may prefer one state for pragmatic reasons and, therefore, wish to analyze the workings only within the one state. Here, John Anderson's (1938, p. 128) notion of a causal field is helpful (see also Mackie 1980, p. 35; Hoover 2001, pp. 41–49). The *causal field* consists of background conditions that, for analytical or pragmatic reasons, we would like to set aside in order to focus on some more salient causal system. We are justified in doing so when, in fact, they do not change or when the changes are causally irrelevant. In terms of representation within the structural account, setting aside causes amounts to fixing certain parameters to constant values. The effect is not unlike Pearl's or Woodward's wiping out of a causal arrow, though somewhat more delicate. The replacement of a parameter by a constant amounts to absorbing that part of the causal mechanism into the functional form that connects the remaining parameters and variables. We might, for instance, wish to conduct our analysis of the *African Queen's* steam engine entirely in the spit-and-polished state, by setting the parameter governing the state of the valve to *clean* and holding it there.

Cartwright's toaster can be treated in the same manner. A parameter might represent the state of the bolt holding the rack to the lever: when it takes the value *tight*, the operation of the toaster is "as advertized"; when *loose*, it is a little wonky; when *missing*, it does not pop up the toast at all. While there are purposes for which only *tight* matters and for which we can treat the bolt parameter as a constant with that value, impounding the state of the bolt to the causal field, it would miss a critical point not to notice that broken mechanisms are mechanisms of the same type as well-functioning mechanisms or that less refined descriptions of mechanisms are special cases of more refined descriptions. Recognizing the first is essential to the repairman; recognizing the second is essential to the design engineer.

The structural account supplemented with the notion of the causal field provides a tool through which different models, different perspectives on phenomena, may be brought into systematic relationship one to another. It also allows us to understand hierarchical relationships among causal systems stressed by Simon (1996) and Wimsatt (2007). While the examples so far involve physical mechanisms, economic examples abound. Cochrane (1998, p. 283) points out in a monetary-policy system similar to (3.13) and (3.14) that α is interpreted as the slope of the aggregate-supply curve and is typically treated by monetary economists as a parameter; yet a body of economic theory and empirical analysis treats α as a

variable, determined by "deeper" parameters (Lucas 1972, 1973; Hoover 1988, Chap. 2). The monetary-policy system impounds these deeper parameters in the causal field.

The monetary-policy rule in (3.13) offers another example. As written, we can analyze the effects of different settings of the parameter λ. We can also consider an "institutional" change in which the rule is altered to depend on different conditioning variables or in different ways. As it stands, these alternatives have been impounded in the causal field. When released and represented in a model in which (3.13) is a special case, we can consider which is the best rule within the now wider class of rules, which contains the current rule as one parameterization (see Woodford 2003 for an extensive discussion of optimal monetary rules).

3.2 Interventions and Identity

The structural account of causal order is similar to Woodward's account in a number of ways. A key difference, however, is that direct cause is not defined with respect to a token-level notion, such as Woodward's come-what-may intervention. Direct cause is expressed instead entirely with respect to the type-level relationship between a privileged set of parameters and a functional relation representing the interrelationship among variables. The structural account, as we showed in the last section, supports a notion of the identity of causal systems: two causal systems are identical if, and only if, they differ at most by their parameterization (i.e., they differ only in the token settings of the parameters). This understanding of causal identity also suggests a different conception of an intervention.

The notion of a parameter developed in the structural account was inspired by Simon's notion of direct control and the notion that the parameter space could be thought of as the loci of direct control. Direct control is virtually indistinguishable from Woodward's notion of intervention. The only question that separates them is direct control of what? For Woodward, it is direct control of a variable; for the structural account, it is direct control of a parameter. But a parameter was defined to be a variable subject to some additional constraints; so the difference seems small. I have no doubt that the experience of manipulation and control are the source of our original intuitions about causal powers and, therefore, are important in the way that we learn about causes and learn to use causal language. Nonetheless, the structural account does not actually *use* the notion of direct control in any physical or metaphysical sense to define causal order. For Woodward an intervention involves a change to the bearer of a variable – a real entity. In contrast, the notion of a parameterization does not require that we change any parameter in a temporal or genetic sense, but merely that we consider different settings of parameters in otherwise causally identical systems.

Woodward accepts that the relevant intervention could be hypothetical and certainly need not be practically implementable (e.g., removing the moon to discover its effect on the tides would be an acceptable, but hypothetical,

intervention). Nonetheless, the counterfactual that is entertained is still a particular token change to a particular entity. In contrast, the structural account is a thoroughly type-level account. The parameter space is the loci of possible interventions; nevertheless, it is the topology that the parameterization imposes on the variables, rather than any – even hypothetical – selection of particular parameter values that defines the causal structure.

The difference between Woodward's account and the structural account is illustrated in comparative static analysis – a technique familiar to economists. For example, we might ask, what is the effect *ceteris paribus* of a higher rate of inflation on the level of prices? The answer given by the quantity theory of money is that the level of prices is lower for the same quantity of money in circulation when their rate of change is higher (Cagan 1956). The experiment cannot be conducted on a single, actual economy since an increase in the rate of inflation necessarily increases the price level. In principle, we could address the question by considering two economies with identical causal structures differing only in the parameterization necessary to produce distinct inflation rates and evaluate their price levels at the time that their money stocks happen to be equal. (This is, of course, a practical impossibility; we do not have a box of causally identical economies to draw from, but difficulty is different from the in-principle impossibility of changing the inflation rate without changing the price level.)[11] Such comparative static analysis may be relevant to actual economies in just those circumstances that some causal channels are so weak that we can neglect them (or impound them in the causal field) or that we can account for them by conditioning. These are exactly the strategies that Cagan (1956) attempts to implement in his famous paper on hyperinflations.

Comparative statics in well-formulated models are examples of a kind of possible-worlds analysis in which the connection to our world – or at least to the model that we take to best represent our world – is substantially more precise than the metrics proposed by Lewis (1973, 1979). Counterfactual analysis and which counterfactuals are sensible to address which particular problems is straightforward in such models (see Hoover 2011).

Something like comparative static analysis is critical to design and engineering. One might, for example, want to understand the difference in behavior of cars of the same model, differing only in having either a four- or a six-cylinder engine.

Motivated in part by his particular conception of an intervention and in part by an essentialist ontology, Woodward rejects the notion that such questions are properly causal with respect to some properties, including race, sex, and species:

> the notion of an intervention will not be well-defined if there is no well-defined notion of changing the values of that variable. Suppose that we introduce a variable "animal" which takes the values {*lizard*, *kitten*, *raven*}... we have no coherent idea of what it is to change a raven into lizard or kitten. Of course, we might keep a raven in a cage and replace it with a

[11] My view here as a shift from my earlier understanding of the causal significance of comparative static analysis (Hoover 2001, p. 102).

> lizard or a kitten, but this is not to change one of these animals into another. What is changed in this case is the content of the cage, not the animals themselves. (Woodward 2003, p. 113)

The issue arises in economic and sociological research when, for example, race discrimination in mortgage applications or sex discrimination in employment is assessed by sending applicants who have been matched as thoroughly as feasible for salient characteristics, differing significantly only in race or sex, through a mortgage qualification or hiring process. With respect to sex discrimination, Woodward (2003, p. 115) rejects the claim "[b]eing female causes one to be discriminated against in hiring and/or salary" as "fundamentally unclear" since "we lack any clear idea of what it would be like to manipulate it." Woodward argues that it is not the applicant's sex but the employer's beliefs about them that is causal.

How are sex and race different from considering the causal outcomes of cars that differ only by their engine type? Consider a coin-sorting machine – coin discrimination being less emotive than sex or race discrimination. Different coins can be placed into the machine and fall into different slots depending on their shapes.

The mechanically relevant description of a coin is as a vector of variables – for example [*diameter*, *thickness*, *weight*]. What else is it to be a coin other than having the right values in such a vector? The social metaphysical answer might be that the essence of being, say, a nickel is to have the imprimatur of the government (to have been dubbed legal tender for $0.05 by the United States Mint) and to have an appropriate standing in the social practice of money. In other contexts, such considerations may be of genuine interest, but they are not *mechanically* salient.

For a causal understanding for purposes of designing, building, operating, maintaining, and repairing the coin sorter, we do not need a penny to change into a nickel. Each penny, nickel, dime, and quarter is just *the coin in the machine* – an instantiation of a vector-valued variable and its causal fate is a realization of the causal process represented by the causal connections among that variable and those that describe the state of the machine.

We need to distinguish between existential and causal identity. Each may be salient in different contexts. The critical question is which context is relevant for what purposes. We have no notion of how to turn a penny into a nickel, but we have a clear notion of how to change a coin from being a nickel to being a penny in the context of the operation of a coin sorter, and it seems perfectly sensible to say that it is "being a nickel" that is a cause of the coin falling into slot 3. Slot 3 being configured in the right way is, of course, another cause, the causes interacting according to the design of the machine.

How is sex discrimination different in principle from coin discrimination? In the hiring process, we can represent people as a vector-valued variable *Applicant* = [*Race, Sex, Age, Employment Status, Wealth,*...], which interacts with variables describing the other factors and processes related to hiring. We do not need to change a particular person from male to female to understand this causal process. It is enough to reparameterize the vector. Much of the effort in conducting discrimination research using such techniques goes in to establishing the relevant causal

identity by closely matching relevant characteristics of different applicants – other than the target characteristic of sex or race. Significant knowledge is obtainable in such ways and there is no need to stigmatize its causal *bona fides*.

Woodward's argument that it is the beliefs of the employer not the being female that is the relevant causal variable does not persuade. The beliefs of employers may be causes of the detailed outcomes of the discrimination – for example, not hiring or paying a lower salary. The relevant question here, however, is what causes those beliefs. The most common reason that we believe someone to be female is that she *is* female. Woodward could object that it is not being female that causes the belief but the appearance of being female. But this is just the analogue of saying that it is not the "moneyness" of the nickel, but its physical characteristics that cause it to fall into slot 3. The physical characteristics *are* the causally relevant ones. What is it to be female in a causally relevant sense? It is to have a sufficient number of stereotypical female characteristics. For purposes of the employment process, a sufficiently plausible transvestite counts as female. Characteristics of actual females determine the female stereotype – that is, why the values of certain variables bundled in a certain way convey the appearance of femaleness. And without the stereotype, variables with those values would not be causally salient. (The same issue arises with the coin sorter: a slug is to a nickel as a transvestite is to a female.)

Some entities may have an essence that cannot be changed while maintaining existential identity. But if, as the structural account would have it, token manipulations are not essential to defining cause, then it is better in these cases to say that causal identity is determined by possession of the explicitly causally relevant characteristics and not by some *sine qua non.* It is also pragmatically superior if we accept the view that knowledge is acquired in a piecemeal fashion. As we have seen, variables such as sex or race are avoided through substantial extra articulation and refinement of causal mechanisms, which may lack clear conceptual or evidential foundations. What, for example, are the detailed mechanisms by which the appearance of being female translates into beliefs and how are they causally relevant? We may not have grounds for knowing such details and we may not need to know them to know what is pragmatically relevant about discrimination. And if I am correct that the coin sorter is really no different in principle than the sex discrimination case, then we would be forced to look for such extra articulation in a vast array of cases. But if the structural account is plausible, we can do happily without it and without the troubling requirement of token intervention.

4 Causation and Representation

The focus of this chapter has been on representing causes. Of course, the ultimate importance of causal analysis is not located in its representation but in discovering what causal relationships actually obtain in the world and using those causes to

control the world for desirable ends – what Cartwright (2007) refers to as "hunting" and "using" causes. Either activity, however, is greatly furthered by a good representation: think of the utility of having a good wanted poster or a good set of blueprints. What I hope to have shown in this chapter is that a relatively straightforward system of causal representation can be useful in understanding and resolving substantive debates in the philosophy of causation, as well as in the actual applications of causal inference and manipulation in economics.

References

Anderson, John. 1938. *The problem of causality*. Reprinted in *Studies in empirical philosophy*. Sydney: Angus Robertson, 1962.

Cagan, Philip. 1956. The monetary dynamics of hyperinflation. In *Studies in the quantity theory of money*, ed. Milton Friedman. Chicago: University of Chicago Press.

Cartwright, Nancy. 1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Cartwright, Nancy. 2007. *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

Cartwright, Nancy, and Martin Jones. 1991. How to hunt quantum causes. *Erkenntnis* 35: 205–231.

Cochrane, John H. 1998. What do the VARs mean? Measuring the output effects of monetary policy. *Journal of Monetary Economics* 41(2): 277–300.

Giere, Ronald N. 2006. *Scientific perspectivism*. Chicago: University of Chicago Press.

Haavelmo, Trgyve. 1944. The probability approach in econometrics. *Econometrica* 12 (supplement): 1–115.

Hamilton, James D. 1995. Rational expectations and the econometric consequences of changes in regime. In *Macroeconometrics: Developments, tensions, and prospects*, ed. Kevin D. Hoover, 325–345. Boston: Kluwer.

Hausman, Daniel M. 1998. *Causal asymmetries*. Cambridge: Cambridge University Press.

Hausman, Daniel M., and James Woodward. 1999. Independence, invariance, and the causal Markov condition. *The British Journal for the Philosophy of Science* 50(4): 521–583.

Hausman, Daniel M., and James Woodward. 2004. Modularity and the causal Markov condition: A restatement. *The British Journal for the Philosophy of Science* 55: 147–161.

Hood, William C., and Tjalling C. Koopmans (eds.). 1953. *Studies in econometric method*, Cowles Commission Monograph 14. New York: Wiley.

Hoover, Kevin D. 1988. *The new classical macroeconomics: A sceptical inquiry*. Oxford: Blackwell.

Hoover, Kevin D. 2001. *Causality in macroeconomics*. Cambridge: Cambridge University Press.

Hoover, Kevin D. 2004. Lost causes. *Journal of the History of Economic Thought* 26(2): 149–164.

Hoover, Kevin D. 2008. Causality in economics and econometrics. In *The New Palgrave dictionary of economics*, 2nd ed, ed. Steven N. Durlauf and Lawrence E. Blume. New York: Palgrave Macmillan.

Hoover, Kevin D. 2009. Milton Friedman's stance: The methodology of causal realism. In *The methodology of positive economics: Milton Friedman's essay fifty years later*, ed. Uskali Mäki, 303–320. Cambridge: Cambridge University Press.

Hoover, Kevin D. 2011. Counterfactuals and causal structure. In *Causality in the sciences*, ed. Phyllis McKay Illari, Federica Russo, and Jon Williamson, 338–360. New York: Oxford University Press.

Hoover, Kevin D. 2012a. Economic theory and causal inference. In *Handbook of the philosophy of economics*, ed. Uskali Mäki. Amsterdam: Elsevier/North-Holland.
Hoover, Kevin D. 2012b. Pragmatism, perspectival realism, and econometrics. In *Economics for real: Uskali Mäki and the place of truth in economics*, ed. Aki Lehtinen, Jaakko Kuorikoski, and Petri Ylikoski. London: Routledge.
Hoover, Kevin D. 2012c. Causal structure and hierarchies of models. In *Studies in History and Philosophy of Science (Part A)* 43(4): 778–786.
Koopmans, Tjalling C. 1950. *Statistical inference in dynamic economic models*, Cowles Commission Monograph 10. New York: Wiley.
Lewis, David. 1973. Causation. *Journal of Philosophy* 70: 556–567.
Lewis, David. 1979. Counterfactual dependence and time's arrow. *Noûs* 13(4): 455–476.
Louçã, Francisco. 2007. *The years of high econometrics: A short history of the generation that reinvented economics*. London: Routledge.
Lucas Jr., Robert E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Lucas Jr., Robert E. 1973. Some international evidence on output-inflation tradeoffs. *American Economics Review* 63: 326–334.
Lucas Jr., Robert E. 1976. *Econometric policy evaluation: A critique*. Reprinted in Lucas, *Studies in business cycle theory*, 104–130. Oxford: Blackwell.
Mackie, John L. 1980. *The cement of the universe: A study in causation*, 2nd ed. Oxford: Clarendon.
Pearl, Judea. 2000. *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Reichenbach, H. 1956. *The direction of time*. Berkeley: University of California Press.
Reiss, Julian. 2007. *Error in economics: The methodology of evidence-based economics*. London: Routledge.
Simon, Herbert A. 1996. *The sciences of the artificial*, 3rd ed. Cambridge, MA: MIT Press.
Simon, Herbert A. 1953. Causal order and identifiability. in Hood and Koopmans (1953). Page numbers refer the reprint in Simon (1957). *Models of man*. New York: Wiley.
Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*, 2nd ed. Cambridge, MA: MIT Press.
Suppes, Patrick. 1970. *A probabilistic theory of causality*. Amsterdam: North-Holland.
Teller, Paul. 2001. Twilight of the perfect model model. *Erkenntnis* 55(3): 393–415.
Wimsatt, William C. 2007. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.
Woodford, Michael. 2003. *Interest and prices: Foundations of a theory of monetary policy*. Princeton: Princeton University Press.
Woodward, James 2003. *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Part II
Models and Representation

Chapter 4
The Regrettable Loss of Mathematical Molding in Econometrics

Marcel Boumans

Abstract Although most accounts on causality discuss the specific role statistics and theory should have, it is taken for granted that they at least have a role in finding causal structures. The role for mathematics is not so obvious. However, before what is called the Probabilistic Revolution in econometrics, identification of causal relations was not a matter of economic-theoretical and statistical significance alone. Mathematical molding was considered as an essential tool in finding significant causal factors. In the 1940s, mathematical molding disappeared in the changeover from methods to specify causal mechanisms of business cycles to methods to identify economic structures, that is, invariant relationships underlying the workings of an economy. Mathematical molding could fulfill its role in modeling business cycle mechanisms because of the assumed close connection between mathematical representations of the business-cycle phenomenon and those of the explanatory mechanism. When the econometric program shifted its focus from mechanisms explaining phenomena to uncovering structural relationships, direct feedback from the phenomenon to the mechanism was lost and the role of mathematical molding ceased to exist.

1 Introduction

Although most accounts on causality discuss the specific role statistics and theory should have, it is taken for granted that they at least have a role in finding causal structures.[1] The role for mathematics is not so obvious. Exemplary for this general

[1] See, for example, the opening sentence of Aldrich's (1989) paper on Autonomy: "Knowledge of structure is valuable and available – but only to those prepared to use both economic theory and statistical analysis."

M. Boumans (✉)
Amsterdam School of Economics, University of Amsterdam, Valckenierstraat 65, 1018 XE Amsterdam, The Netherlands

Erasmus Institute for Philosophy and Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands
e-mail: m.j.boumans@uva.nl

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_4,

view of the role of mathematics is the Cowles Commission econometric approach. The Cowles Commission view (see, e.g., Christ 1994) was that to understand a particular aspect of economic behavior, it is necessary to have a system of descriptive equations. These equations should contain relevant observable variables, be of known form (preferably linear), and have estimatable coefficients. However, "little attention was given to how to choose the variables and the form of the equations; it was thought that economic theory would provide this information in each case" (Christ 1994, p. 33). This position was explicitly expressed by Tjalling Koopmans, director of the Cowles Commission, in a paper jointly written with Herman Rubin and Roy B. Leipnik, "Measuring the Equation System of Dynamic Economics."

> The analysis and explanation of economic fluctuations has been greatly advanced by the study of systems of equations connecting economic variables. The construction of such a system is a task in which economic theory and statistical method combine. Broadly speaking, considerations both of economic theory and of statistical availability determine the choice of the variables. (Koopmans et al. 1950, p. 54)

However, before what is called the Probabilistic Revolution in econometrics, identification of causal relations was not a matter of economic-theoretical and statistical significance alone. Mathematical molding was considered as an essential tool in finding significant causal factors. According to Ragnar Frisch's (1933a) original econometric ideal, all three "viewpoints," economic theory, statistics, and mathematics, were necessary, but not by themselves sufficient: "It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics" (Frisch 1933a, p. 2).

This founding ideal of the Econometric Society, that is, the union of mathematics, economics, and statistics, however, was lost in later econometric-modeling practices. In the 1940s, mathematical molding disappeared from the econometric scene, as Mary Morgan describes in her *History of Econometric Ideas* (1990, p. 264):

> Between the 1920s and the 1940s, the tools of mathematics and statistics were indeed used in a productive and complementary union to forge the essential ideas of the econometric approach. But the changing nature of the econometric enterprise in the 1940s caused a return to the division of labour favoured in the late nineteenth century, with mathematical economists working on theory building and econometricians concerned with statistical work.

Mathematical molding disappeared in the changeover from methods to specify causal mechanisms of business cycles to methods to identify economic structures, that is, invariant relationships underlying the workings of an economy. Mathematical molding could fulfill its role in modeling business-cycle mechanisms because of the assumed close connection between mathematical representations of the business-cycle phenomenon and those of the explanatory mechanism. When the econometric program shifted its focus from mechanisms explaining phenomena to uncovering structural relationships, direct feedback from the phenomenon to the mechanism was lost, and the role of mathematical molding ceased to exist.

The works of the Jan Tinbergen, discussed in Sect. 2, show how mathematics was and could be used for identification purposes. The method Tinbergen employed to arrive at a causal explanation of the business-cycle phenomenon started with a priori economic-theoretical considerations about which explanatory variables should be included. Some of the explanatory variables appeared as differential or integral terms in the model equations. The equations were chosen to be linear, and the values of the parameters were found by multiple regression analysis. Statistical tests of significance were applied to measure the accuracy of these results. And, moreover, the parameter values found for the causal relations were adjusted to make sure that the model yields a cyclic movement with characteristics in accordance with those of the actual business cycle. As a result of this latter assessment (which will be called "tuning"), it appeared that integral terms were not of any significance and therefore could be neglected. Differentials were approximated by differences. So, after starting with using mixed differential-difference-integral equations, Tinbergen ended up with representations of the business-cycle mechanism with only difference equations.

In response to Tinbergen's reports on this method, Frisch (1995), discussed in Sect. 3, showed that the initial close relationship between the specific mathematical representation of the business cycle and the mathematical representation of its mechanism was lost in the transformation to difference equations. As a result, it was no longer possible to identify all relevant causal factors. "Passive observation" alone is not sufficient to detect them; statistics alone cannot reveal inactive but potential factors. Without any feedback from the phenomena, we have to rely on economic theory to provide us with a complete list of factors. A similar critique was brought forward by John Maynard Keynes. Although unjustly addressed to "Professor Tinbergen's Method," it certainly applies to the later Cowles Commission approach.

> Am I right in thinking that the method of multiple correlation analysis essentially depends on the economist having furnished, not merely a list of the significant causes, which is correct so far as it goes, but a *complete* list? For example, suppose three factors are taken into account, it is not enough that these should be in fact *veræ causæ*; there must be no other significant factor. If there is a further factor, not taken account of, then the method is not able to discover the relative quantitative importance of the first three. If so, this means that the method is only applicable where the economist is able to provide beforehand a correct and indubitably complete analysis of the significant factors. The method is one neither of discovery nor of criticism. It is a means of giving quantitative precision to what, in qualitative terms, we know already as the result of a complete theoretical analysis. (Keynes 1939, p. 560)

In other words, taking a strong apriorist position means that econometrics becomes a method not of testing or of discovery, but of measurement.

Haavelmo (1944), discussed in Sect. 4, discussed the problem of finding a complete list of causal factors under the heading of the "problem of autonomy." However, the problem of autonomy was broader than this; it also covered the problem of invariance. This latter issue concerns the identification of the relationships between the causal factors that remain unaffected by changes

elsewhere in the system. The problem of listing the causal factors and the problem of invariance are closely related in which the requirements of economic-theoretical as well as statistical and mathematical significance all have equal weights.

2 Tinbergen's Business-Cycle Schemes

In the 1920s, empirical business-cycle research mainly consisted of constructing so-called barometers to forecast business cycles. That is to say, their research focused on investigating whether certain economic time series were correlated. If there is a lag between correlated time series, then it is possible to forecast the course of one time series with the aid of the other.

The Harvard Committee on Economic Research (run by Charles J. Bullock, Warren M. Persons, and William L. Crum) owed its international fame to such a "barometer" based on three indices of the business cycle, the so-called *A–B–C* curves. These three indices represented "speculation" (*A*), "business" (*B*), and "money" (*C*) and were lag correlated. *B* lagged about 6 months behind *A*, and *C* lagged about 4 months behind *B*. Therefore, *A* could forecast *B* and both *A* and *B* could forecast *C*.

Tinbergen opposed the nontheoretical character of the Harvard barometer. His very first scientific publication, "Over de Mathematies-Statistiese Methoden voor Konjunktuuronderzoek" (1927; On Mathematical-Statistical Methods of Business-cycle Research), was a review of this kind of business-cycle research. It criticized the Harvard approach for not being based on any kind of causal theory. Moreover, Bullock et al. (1927, p. 79) had admitted that their method was not based on any theory whatsoever; on the contrary, the curves were "derived solely from observation of the facts."

> Causal relations have, indeed, received increasing attention from us; but no theory of causation or of time relation between cause and effect ever entered into the construction of the index. (Bullock et al. 1927, p. 79)

In addition, they observed "how foreign to actual experience are fixed mechanical, or exact mathematical, relationships in the economic world" (p. 79).

Tinbergen (1927) claimed that the aim of correlation analysis should ultimately be the recovery of causal connections, such as Karl G. Karsten's "theory of quadrature" had suggested. Karsten's (1926) had shown the existence of two "cumulative relations" between the three Harvard barometer indices, which he interpreted as causal relationships. In the first place, he found, using correlation analysis, that the cumulative values of the Harvard *B*-index parallel those of the Harvard *A*-index, with a lag of 3 months:

$$\sum_{i=1}^{t} B_i = A_{t+3} \tag{4.1}$$

Second, he found the empirical relationship that the C-index was a cumulative sum of both the A and B indices:

$$\sum_{i=1}^{t}\left(\frac{1}{4}A_i+\frac{3}{4}B_i\right)=C_t \tag{4.2}$$

Thus, according to Karsten (1926, p. 417), the B-index was the "generating force" of the three; the other two indices depended upon, and were derived from, changes in the business index.

Equations (4.1) and (4.2) express cumulative relations of discrete processes. For continuous processes, cumulative relations can be expressed by means of integrals,[2]

$$\int_0^t B(\tau)\mathrm{d}\tau=A(t+3) \tag{4.3}$$

or by differentials:

$$B(t)=\frac{\mathrm{d}A(t+3)}{\mathrm{d}t}=\dot{A}(t+3) \tag{4.4}$$

In classical mechanics, there is a close connection between the calculus of variations and cause-and-effect relations. It is because of this connection that Karsten wanted to apply the "theory of quadrature" to investigate the kind of relations that exists between economic quantities. When a cause-and-effect relation exists between two phenomena, then according to the quadrature theory, one phenomenon is expected to be cumulatively affected by the other:

> In the calculus such relations are familiar in the form of integrals and derivatives, and although these functions are purely mathematical, they are useful to describe the behavior of related forces in the physical sciences. It is the quadrature theory that economic data or statistics betray the same relationships when similarly treated, and that when this is the case, the economic forces or phenomena measured by statistics may be said to be in quadrature and a real relation is strongly suggested. (Karsten 1924, p. 14)

Tinbergen found these cumulative relations exemplary for the kind of causal relation one could expect to find in business-cycle research. It was the application of this connection between the calculus and causal relationships that made Karsten's approach so appealing to Tinbergen.

Tinbergen was looking for causal explanations of business cycles, but economic theory did not provide the appropriate mechanisms. On the one hand, business cycles were explained by exogenous influences; on the other hand, each cycle was examined and explained individually, or, worse still, each phase of a cycle was

[2] This explains the name "quadrature theory." Quadrature stands for the process of determining the area of a plane geometric figure by dividing it into a collection of shapes of known area (usually rectangles) and then finding the sum of these areas. The integral denotes this process for infinitesimal rectangles.

explained separately. However, Albert Aftalion's (1927) "Theory of Economic Cycles Based on the Capitalistic Technique of Production" was an exception, on which Tinbergen commented:

> An economic dynamics could be constructed based on the [lag] relation between economic quantities, which results in the derivation of perfect cyclic oscillations of an economic system. This is the mathematical interpretation of Aftalion's crisis theory. I mention this theory in particular because it explains most clearly how the relations considered here can happen, in that every cycle already contains the seed for the next cycle and thus real periodicity occurs. (Tinbergen 1927, p. 715[3])

Aftalion's thesis was "that the chief responsibility for cyclical fluctuations should be assigned to one of the characteristics of modern industrial technique, namely, the long period required for the production of fixed capital" (Aftalion 1927, p. 165). For producers, the value of a product depends on the price it is expected to fetch; that is to say, their value depends on the forecast of future prices. Aftalion assumed that the expectations of these future prices are, alternately, either too optimistic or too pessimistic: "the rhythm is a consequence of the long delay which often separates the moment when the production of goods is decided upon and a forecast is made from the moment when the manufacture is terminated, and the forecast is replaced by reality" (p. 165). Producers forecast future prices on the basis of present prices and the present state of demand. "That is the source of their errors. In modern capitalistic technique the actual state of demand and prices is a bad index of future demand and prices, because of the long interval which separates the moment when new constructions are undertaken from that when they satisfy the demand" (p. 166).

In a paper, "Opmerkingen over Ruilteorie" (Observations on Exchange Theory) published in 1928, Tinbergen constructed a numerical example demonstrating how a delayed adjustment of supply to price would generate fluctuations about equilibrium over time. Shortly after this, he stumbled across an empirical example of this numerical construction in a pork-market study by Arthur Hanau (1928) (Tinbergen 1928, p. 548n; see also Magnus and Morgan 1987, p. 120). According to Tinbergen, this scheme of delayed supply adjustment to price could be extended by taking into account expectations based on observed past fluctuations or by attributing a delay to demand. "All these assumptions lead to the same kind of results, of which the essence ... consists in the explanation of cyclic motion by the economic mechanism itself" (Tinbergen 1928, p. 546[3]).

At the first European meeting of the Econometric Society in 1931, Tinbergen (1933a) had a number of mathematical formalizations of an endogenous business-cycle mechanism to offer for consideration. Hanau's (1928) research into the pork-market, "le cas le plus simple," served as point of departure:

Scheme I	Supply:	$A_0 + A_1 p(t - \theta)$
	Demand:	$B_0 - B_1 p(t)$

where A_0, A_1, B_0, and B_1 are positive constants and $p(t)$ the deviation from the equilibrium price P at time t. θ was the time needed to produce the relevant

[3] Translated by the author.

commodity. The mechanism represented by this scheme generated a cycle with a period equal to 2θ. This scheme, known as the cobweb mechanism because of the likeness between its graphical representation and a cobweb, was the simplest explanation of an economic cycle and a mathematical generalization of Tinbergen's earlier numerical example.

However, the aim was to find mechanisms that could explain the so-called Juglars. These were business cycles with a cycle period of about 6–10 years. Therefore, *Scheme I* that implied a production time of 3–5 years is unrealistic for most production processes. To arrive at a more realistic representation of business cycles, Tinbergen examined more advanced schemes to see what influence each added "complication" could have on the length of the cycle period.

In a second scheme, he introduced "demande spéculative." There was some empirical evidence that demand could also be influenced by price changes, $\dot{p}(t)$, as was seen in the wholesale lumber trade, or corn speculation.

Scheme II Supply: $A_0 + A_1\, p(t - \theta)$
Demand: $B_0 - B_1 p(t) + B_2 \dot{p}(t)$

In *Scheme II*, the period of the solution (T) lies between: ${}^4/_3\,\theta < T < 2\theta$. So, the introduction of a differential shortens the period of the business cycle with respect to the production lag. In other words, *Scheme II* could not be considered as a possible explanation for the Juglar; it assumes an even longer production time.

Another way of advancing *Scheme I* was to introduce purchasing power into the demand function. First, Tinbergen considered constant purchasing power, C.

Scheme III Supply: $A_0 + A_1 p(t - \theta)$
Demand: $\dfrac{C}{P + p(t)}$

The solution of this scheme had a period length equal to 2θ. So, constant purchasing power did not increase the cycle period. Next, he assumed purchasing power dependent on economic activity defined as the numbers of workers employed during the production process:

$$N(t) = \alpha \int_{t-\theta}^{t} [A_0 + A_1 p(\tau)]\, \mathrm{d}\tau \tag{4.5}$$

If wages are constant and equal to S, then total purchasing power equals SN, and the scheme becomes:

Scheme IV Supply: $A_0 + A_1\, p(t - \theta)$
Demand: $\dfrac{S\alpha \int_{t-\theta}^{t} [A_0 + A_1 p(\tau)\mathrm{d}\tau]}{P + p(t)}$

For this scheme, the cycle period was equal to 2.7θ. Thus, by assuming that purchasing power is dependent on economic activity, Tinbergen was able to extend the period compared with the production lag and thus arrived at a more realistic business-cycle mechanism.

In his search for appropriate business-cycle mechanisms, Tinbergen (1931) found an even better example than the "pork cycle": the shipbuilding cycle. This mechanism, a combined lag and cumulative relation, showed how a lag of 2 years could generate a cycle of 8 years:

$$\dot{X}(t) = -aX(t-\theta) \tag{4.6}$$

where X represents world tonnage and θ the average needed time to build a ship, approximately 2 years. The parameter a was a constant value between ½ and 1. The cycle generated by this mechanism has a period equal to $4\theta = 8$ years.

In a survey on "quantitative business-cycle theory," Tinbergen (1935) outlined systematically the criteria for an appropriate business-cycle theory: "The aim of business cycle theory is to explain certain movements of economic variables. Therefore, the basic question to be answered is in what ways movements of variables may be generated" (p. 241). The core of the business-cycle theory was a "mechanism" that he defined as a "system of relations existing between the variables; at least one of these relations must be dynamic. This system of relations defines the structure of the economic community to be considered in our theory. Such a mechanism may perform certain kinds of swinging movements that are characteristic of the system as such" (pp. 241–2). A mechanism, according to Tinbergen, is a specific set of structural relations that together explain the business cycle. Tinbergen emphasized the distinction between the mathematical form and the economic meaning of the equations:

> The mathematical form determines the nature of the possible movements, the economic sense being of no importance here. Thus, two different economic systems obeying, however, the same types of equations may show exactly the same movements. But, it is evident that for all other questions the economic significance of the equations is of first importance and no theory can be accepted whose economic significance is not clear. (Tinbergen 1935, p. 242)

Mathematical molding was an essential element of Tinbergen's business-cycle research in the 1930s. Economic theories did not contain any guidelines that could lead to an appropriate formalism. They either were verbal accounts or, if mathematical, only gave descriptions of *static* systems. Mathematical molding was a trial-and-error process that started with the assumption of a production lag. As Hanau (1928) showed empirically and Aftalion (1927) theoretically, lags generate endogenous fluctuations. However, basing dynamics on a production lag alone has several disadvantages. In the first place, as discussed above, to explain a Juglar, the assumed production time would have to be far too long. This was the reason why Tinbergen introduced various "complications" into his dynamic schemes. In the second place, the disadvantage of postulating lags is that they must be stated in advance and have a fixed length. "This has been repeatedly felt as a too rigid

representation of reality" (Tinbergen 1933b, p. 13[3]). However, beside lag relations other dynamic relations are possible, namely, those containing differentials and integrals. From physics, Tinbergen knew that second-order differential equations can generate cycles. For example, differentiating (with respect to time) an equation containing a differential and an integral term leads to the equation of the harmonic oscillator.

$$a\dot{y}(t) + by(t) + c\int_0^t y(\tau)\mathrm{d}\tau = 0 \rightarrow a\ddot{y}(t) + b\dot{y}(t) + cy(t) = 0 \tag{4.7}$$

An advantage of differential equations is that differentials refer to very small time intervals. Note that $\dot{y} = \mathrm{d}y/\mathrm{d}t$, where $\mathrm{d}t$ can be approximated by a very small difference in time Δt. So that:

$$\dot{y} \approx \frac{y(t) - y(t - \Delta t)}{\Delta t} \tag{4.8}$$

> Considering the shorter time many production processes need nowadays, the appearance of only direct affective causes can be called a realistic feature in view of this. Thus, what really matters is the question just posed: can quantities with an integral character and a differential character, respectively, be found and do these quantities play an important role in the business cycle? (Tinbergen 1933b, pp. 14–15)

At a meeting of the Econometric Society in Leiden in 1933, Tinbergen raised this question most explicitly: "Is the theory of harmonic oscillation useful in the study of business cycles?" He proposed to start "from the mathematical nature of harmonic oscillations and seeking among the main economic relations those likely to fit into the harmonic pattern" (Marschak 1934, p. 188). Accordingly, he marshaled economic relations into two groups: (1) "differential phenomena," that is, functions of the rate of price change, $\dot{p}(t)$, and (2) "integral phenomena," that is, functions of $\int p\,\mathrm{d}t$. Statistical tests, however, showed him not to give too much credit to most of the phenomena of group (2), because the correlations he had hitherto found were too small (p. 188).

In his 1935 survey, Tinbergen discussed this issue again. To make "closer approximations to reality" (p. 277), differentials, $\dot{p}(t)$, and integrals, $\int p\,\mathrm{d}t$, were added to the lag schemes. Thus, in general, the reduced form equation of a business-cycle scheme would have the following shape:

$$\sum_1^n a_i p(t - t_i) + \sum_1^n b_i \dot{p}(t - t_i) + \sum_1^n c_i \int_0^{t-t_i} p(\tau)\,\mathrm{d}\tau = 0 \tag{4.9}$$

The requirement was that the parameters satisfy the "wave condition" and the "long wave condition." The "wave condition" required that the solution to the above reduced form equation is a sine function, $p(t) = C\lambda^t \sin(\omega t)$, so that the time shape of $p(t)$ is cyclic. The "long wave condition" prescribed that the cycle period should be long compared with the "time units" and that the cycle should not differ "too much from an undamped [*sic*] one" (p. 280). According to Tinbergen, "these conditions will be a guide in a statistical test of the different schemes as to their accord with reality" (p. 280). As a first approximation to these conditions, Tinbergen put $\lambda = 1$ and $\omega = 0$. Then the period of the cycle, $2\pi/\omega$, goes to infinity. Both conditions taken together implied that:

$$\sum_{1}^{n} c_i = 0 \tag{4.10}$$

In other words, mechanisms "only then lead to long, not too much damped waves when the integral terms are of small importance" (p. 281).

Tinbergen considered several mechanisms for their ability to explain the business cycle. The wave conditions defined the restrictions on the parameter values. But to find out whether these possible mechanisms "can explain real business cycles and which of them resembles reality" (p. 281), statistical verification, however, was the necessary next step in the analysis.

Tinbergen's research program in the first half of the 1930s can be briefly characterized as a combination of two methods, mathematical molding and statistical verification. Mathematical molding generated potential business-cycle mechanisms, which had to be identified empirically. But also in his subsequent work in the 1930s, when he built the two very first macro-econometric models, mathematical molding was part of the modeling process, although less prominent. The first macro-econometric model was his 1936 model of the Dutch economy. The second was developed when he was commissioned by the League of Nations to undertake statistical tests of the business-cycle theories, published in a two-volume work, *Statistical Testing of Business-Cycle Theories* (1939a, b). The first volume contained an explanation of a method of econometric testing and a demonstration, using three case studies, of what could be achieved. The second volume contained a model of the United States economy.

The procedure he employed to test existing business-cycle theories consisted of two stages. Firstly, the variables that a given theory provides must be tested by multiple regression analysis, and, secondly, the system of numerical values found for the causal relations must be tested to see whether it really yields a cyclic movement when used in the reduced form equation.

Mathematical arguments played a crucial role in the assessment of whether integrals should be built into the model or omitted. This assessment was rather similar to his 1935 discussion of whether integrals should be part of a business-cycle mechanism or not. It appeared that the integrals, now called "cumulants," in

the reduced form equation influenced the possible movements of the system in two ways:

(i) The period and degree of damping of the cyclical movement are to some extent affected by the presence of such terms.
(ii) Besides that, the cumulants introduce an additional root into the characteristic equation, which is real and positive, giving rise to a one-sided movement. This movement is explosive (away from the equilibrium situation) if the algebraic sum of all coefficients of cumulation terms in the final equation [reduced form equation] is positive; the movement is damped (gradual approach of the equilibrium situation) if that sum is negative. (Tinbergen 1939b, p. 147)

However, the cumulants appearing in the reduced form equation were "not all, and perhaps not even the most important of the cumulants to which the economic mechanism gives rise in reality" (p. 149). In many other cases, cumulations could not be distinguished from trends and so remained "hidden." A rough estimate of the possible effects of cumulants (including the hidden ones) showed that they would change the dampening factor at the most by ± 0.05. Because of the hidden cumulants, the sign of the sum of the coefficients for the cumulants could not be determined, but Tinbergen assumed that the positive real root lay somewhere between 0.75 and 1.25, which leaves the possibility of either a dampened or an explosive one-sided movement. The latter possibility had to be rejected for not being in accordance with movements observed in reality. The influence of the former on the cyclical movement would only be moderate. "To sum up, on the ground of their small influence under (i) and our ignorance of their effect under (ii), it seemed both advisable and justified to keep all terms containing cumulants out of the elimination process" (p. 149). So they will not show up in the reduced form equation. So, the result was consistent with his earlier result in his 1935 survey paper that the sum of the coefficients of the integrals should be small (Eq. 4.10) to satisfy the wave conditions.

Like the integrals, nor did the two macro-econometric models contain differentials, but for another reason. Tinbergen had changed his view on the meaning and role of lags in the mathematical relations. In his earlier business-cycle schemes, lags meant production lags and referred to time intervals of about 1–2 years. One of the main reasons for introducing differentials was that they represented more immediate reactions. But in the later macro-econometric models, lags did not have this specific economic meaning anymore; they came to indicate time units of, for example, 1 month. If time lags are time units, $\Delta t = 1$, differentials can be approximated by differences, cf. Eq. (4.8): $\dot{y} \approx y(t) - y(t-1)$.

3 Frisch's Memorandum

Although both volumes of Tinbergen's *Statistical Testing of Business Cycle Theories* were officially published by the League of Nations in 1939, copies of Tinbergen's research were circulated in advance in 1938. Ragnar Frisch wrote a memorandum,

Statistical versus Theoretical Relations in Economic Macrodynamics, to assess Tinbergen's work.

> The present memorandum does not discuss details of the various equations which Tinbergen has obtained and whose coefficients he has determined statistically. My main concern has been to discuss what equations of this type really *mean*, and to what extent they can be looked upon as 'A Statistical Test of Business Cycle Theories'. (Frisch 1995, p. 407)

The memorandum discussed two problems. The first problem was the question "what sorts of equations it is possible to determine from the knowledge of the time shapes that are actually produced" (p. 416). The answer to this question was that only the so-called coflux equations were discoverable. These coflux equations were defined algebraically by the set of functions that forms the actual solution of the complete system, including those determined by the initial conditions. The second and deeper problem was that these coflux relations may not come near to resembling the more "fundamental" equations that form the "essence of theory," the so-called autonomous equations (p. 417). Frisch argued that "it is *only coflux* relations that are determined by Tinbergen, and the lack of agreement between these equations and those of pure theory cannot be taken as a refutation of the latter" (p. 419).

Frisch's analysis of the first problem, here labeled as the "identification problem,"[4] stimulated various members of the Cowles Commission to work on identification in the 1940s (Hendry and Morgan 1995, p. 57). The second problem, the problem of autonomy, was crucial in the development of the concept of structural equations (see Aldrich 1989). Although both problems are closely related, solving the first does not imply a solution to the second. While the identification problem is a mathematical problem, the autonomy problem remains basically an empirical problem.

To understand the nature of both problems, they will be introduced according to their original treatment in Frisch's memorandum. We start with the identification problem. This problem deals with the relation between the "form" of the equations representing the assumed relations between the economic variables and the "time shape" of these variables.

Frisch defined the form of a difference equation,

$$\sum_{i\theta} a_{i\theta} x_i(t-\theta) = 0 \tag{4.11}$$

as the $i\theta$ range of the summation that determines the terms involved in the equation. The time shape of a variable is the sum of the exponentials that make up this variable,

[4] This was not Frisch's terminology, but Koopmans'. Aldrich (1994) gives an account of the development of the identification theory from Frisch to Koopmans by focusing on Haavelmo (1944), including a discussion of the change in terminology.

$$x_i(t) = \sum_{k=1}^{n} C_{ik}\, e^{\gamma_i t} \tag{4.12}$$

In the memorandum, the identification problem was phrased in terms of "reducibility" and "irreducibility" and was linked to the time shapes of the variables.

> It is clear that the property of irreducibility must be important when we are studying the nature of those equations that can be determined from the knowledge of the time shapes of the functions that are to satisfy the equations. (Frisch [1938] 1995, p. 413)

The (ir)reducibility of an equation was defined with respect to a set of functions. An irreducible equation of the form (p. 11) is "one whose coefficients are *uniquely determined* and allow of no degree of freedom *if the equation is to be satisfied by this set of functions* (apart from the arbitrary factor of proportionality which is always present in the case of a homogeneous equation)" (p. 413).

By inserting the function $x_i(t)$ defined in formula (4.12) into Eq. (4.11), one can derive algebraically the following rule[5]:

> Rule about reducibility: If the functions with respect to which reducibility is defined are made up of n exponential components . . ., the equation is certainly reducible – and hence its coefficients are affected in a more or less arbitrary manner – if it contains more than $n + 1$ terms. And it may be reducible even if it contains $n + 1$ terms or less. (Frisch 1995, p. 414)

In other words, only equations that contain at most $n + 1$ terms may be irreducible – uniquely identified – with respect to the time shape of a variable. For example, if the time shape has the form of a (dampened, undampened, or antidampened) sine function, then it is equivalent to a combination of two exponential components and therefore cannot identify an equation with more than three terms.

However, the time shapes of the variables do not satisfy just one equation but form the actual solution of the complete system, including those determined by the initial conditions. Frisch called an equation that is identified by the time shape of this actual solution a "coflux" equation. The other equations were called "superflux" equations. The word "flux" suggested that both kinds of equations were defined with respect to the time shape actually possessed by the phenomena. Thus, only "coflux equations and no other equations are discoverable from the knowledge of the time shapes of the functions that form the actual solution" (p. 416).

> This is the nature of *passive observations*, where the investigator is restricted to observing what happens *when all equations in a large determinate system are actually fulfilled simultaneously*. The very fact that these equations are fulfilled prevents the observer from being able to discover them, unless they happen to be coflux equations. (Frisch 1995, p. 416)

Should one bother about these other equations that are not discoverable through passive observations? Frisch's answer was yes; the other equations, the superflux equations, are well worth knowing because they have a higher degree of

[5] Boumans (1995), which discusses the more technical details of Frisch's memorandum, also provides the derivation of this rule.

"autonomy." These were the equations that "maintained unaltered while other features of the structure were changed" (p. 417).

> The higher this degree of autonomy, the more *fundamental* is the equation, the deeper is the insight which it gives us into the way in which the system functions, in short, the nearer it comes to being a *real explanation*. Such relations form the essence of 'theory'. (Frisch 1995, p. 417)

Unfortunately, autonomy is "not like the irreducibility a mathematical property of a closed system . . . but is built on some sort of knowledge outside this system" (p. 416). Passive observation only lead to coflux equations, and generally spoken these relations are far from able to give information about the autonomous structural relations. Therefore, it is necessary to use active observation, namely, experimentation, as Frisch recommended.

In his memorandum of 1938, the concept of autonomy was not further explicated. Ten years later, Frisch gave a more explicit description of what he meant by the idea of autonomy:

> Take any equation and ask the question: is the technical and institutional setting which surrounds it and the behaviour of the individuals involved such that this particular equation will *hold good* even though other equations involving the same variables are destroyed through technical, institutional or behaviouristic changes or through the fixation of some specific variables in the system, for instance through a specific economic measure. This, it seems, is the only way in which it is possible to define a 'causal' relation as distinguished from an incidental covariation between economic magnitudes. (Frisch 1948, pp. 368–369)

As we have seen above, Tinbergen used the characteristics of the business cycle to acquire information about the causal structure: tests of mathematical significance were used to infer the shape of the equations of the mechanism plus the relevant causal factors. So, an essential part of the model-building process is mathematical molding: a mathematical formalism is sought that is able to generate the relevant characteristics of the phenomena that should be explained or described. Next, the parameters are quantified in such a way that the model precisely picks out these characteristics. This latter stage is called tuning. Because, in the model-building process, mathematical molding and testing for mathematical significance are both sides of the same coin in the model-building process, justification is built-in. In Boumans (1999), where this argument is developed, three examples of models are discussed of which two were built in the same period as Tinbergen's modeling work: Kalecki's (1935) and Frisch's (1933b) business-cycle model. Kalecki tuned his parameters such that his model generated a maintained cycle with a period of 10 years. Frisch tuned the parameters such that his model generated three damped cycles of which two had a period in accordance with the observed cycle periods.

Frisch's memorandum, however, showed that this feedback from the shape of a business cycle to its generating mechanisms was cut off when the mechanism was represented by difference equations instead of (mixed difference-)differential equations. The direct and close relationship between cyclical behavior and differential equations does not exist for difference equations.

This practice of mathematical molding was also criticized by Trygve Haavelmo, a student of Frisch, in his (1940) paper "The Inadequacy of Testing Dynamic Theory by Comparing Theoretical Solutions and Observed Cycles." On the basis of an example, Haavelmo demonstrated that "'correction' of *the form of* a priori *theory* by pure inspection of the *apparent shape* of time series is a very dangerous proceeding and may lead to spurious 'explanations'" (p. 321). The example he gave showed that when an apparent trend, which was "not strongly justified on a priori reasons," is built into the model, things are often assumed to be structural, whereas they are merely the effect of cumulation of random events and thus, in fact, are spurious.

Haavelmo's warning against building time shapes of the variables into the model was one of the arguments for cutting off the empirical feedback from the phenomenon in question to modeling its causal mechanism. The paper was the basis for his later paper on autonomy. In it he showed that when apparent shapes of time series, like temporary trends, are confused with structure, the real explanation for apparent changes in structure is in fact "the disappearance of spurious elements introduced in our theory by the trend fitting" (p. 321). It should, however, be noted that Haavelmo only discussed the danger of building in temporary appearances which are mistaken as steady characteristics of the phenomena or, in other words, as stylized facts about the phenomena. It would take 30 years before the strategy of using facts about a phenomenon to assess parameter values remerged under the new name "calibration." This strategy could only flourish once the high-days of the Cowles Commission approach were over.

4 Haavelmo's Probability Approach

Haavelmo's (1944) *The Probability Approach in Econometrics* echoed Frisch's memorandum in many respects, in particular its terminology, but it is important to be aware of the change in scope. There was not only a shift from linear to nonlinear equations, and the concomitant change in mathematical technique from linear algebra to implicit function theory, but the point of departure also differed (see for this Aldrich 1994, pp. 205–206). Haavelmo envisaged a situation in which the form of the equations is given by the relevant economic theory and the unknowns are the values of the economic structural parameters, while to Frisch all that is given is the possibility there are one or more linear relations between the variables. Nevertheless, Haavelmo also distinguished between the identification problem and the autonomy problem.

The problem of identification was that "one or more of the parameters to be estimated might, in fact, be *arbitrary* with respect to the *system* of equations" (p. 84). This problem came down to a study of the properties of the joint probability distribution of the random (observable) variables in a stochastic equation system. Within this framework, two "fundamental" problems could be formulated, namely,

the "problem of arbitrary parameters" and the problem of "best estimates." The first problem was that if two stochastic equation systems lead to the same joint probability law of the observable random variables, we cannot distinguish between them on the basis of observations (see Haavelmo 1944, p. 88 and p. 91). The second problem was how to find the best estimate for the parameters given a specific sample, in other words a straightforward statistical problem. The first problem was considered a problem of "pure mathematics." "This problem, however, is of particular significance in the field of econometrics, and relevant to the very construction of economic models, and besides, this particular mathematical problem does not seem to have attracted the interest of mathematicians" (p. 92).

The "problem of arbitrary parameters" was described using Frisch's term "reducibility," but with a slightly different meaning. Reducibility was not defined with respect to linear difference equations but to the more general functional equations and was not linked to the exponentials satisfying these equations but to general functions of the parameters. As a result, reducibility was now defined in terms of whether or not the partial derivatives of the functional equations with respect to the parameters were linear dependent.

Thus, while the dual problem of estimation could be tackled in a mathematical and statistical way, the problem of autonomy remained, as in Frisch's memorandum, "a matter of intuition and factual knowledge; it is an art" (p. 29). The problem of autonomy was worded as the problem of "judging the degree of persistence over time of relations between economic variables" or, more general, "whether or not we might hope to find elements of invariance in economic life, upon which to establish permanent 'laws'" (p. 13). This problem is caused by the fact that real economic phenomena cannot be "artificially isolated from 'other influences'" (p. 14). We have to deal with passive observations, and these are:

> influenced by a great many factors not accounted for in theory; in other words, the difficulties of fulfilling the condition 'Other things being equal'. But this is a problem common to all practical observations and measurements; it is in point of principle, not a particular defect of economic time series. If we cannot clear the data of such 'other influences', we have to try to introduce these influences in the theory, in order to bring about more agreement between theory and facts. Also, it might be that the data, as given by economic time series, are restricted by a *whole system* of relations, such that the series do not display enough variations to verify each relation separately. (Haavelmo 1944, p. 18)

Let y denote an economic variable, the observed values of which may be considered as results of planned economic decisions taken by individuals, firms, etc. And let us start from the assumption that the variable y is influenced by a number of causal factors, $x_1, x_2, \ldots$.

> Our hope in economic theory and research is that it may be possible to establish constant and relatively *simple* relations between dependent variables, y (of the type described above), and a relatively *small* number of independent variables, x. In other words, we hope that, for each variable, y, to be 'explained', there is a relatively small number of explaining factors the variations of which are practically decisive in determining the variations of y. (Haavelmo 1944, pp. 22–23)

Haavelmo distinguished between two different notions of "influence," namely, "potential influence" and "factual influence." Let y be a theoretical variable defined as a function of n independent "causal" variables $x_1, x_2, \ldots, x_n$:

$$y = F(x_1, \ldots, x_n) \tag{4.13}$$

Both notions of influences can be clarified by the following equation:

$$\Delta y = F_1 \Delta x_1 + \cdots + F_n \Delta x_n \tag{4.14}$$

The deltas, Δ, indicate a change in magnitude. The terms F_i indicate how much y will change due to a change in magnitude of factor x_i.

Then "potential influence" of the factor x_i upon y can be represented by F_i. Thus, "for a given system of displacements $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$, the potential influences are, clearly, formal properties of the function F" (pp. 23–24). The "factual influence" upon y of the variable x_i can be represented by $F_i \Delta x_i$.[6]

According to Haavelmo, this distinction between potential and factual influence was fundamental.

> For, if we are trying to explain a certain observable variable, y, by a system of causal factors, there is, in general, no limit to the number of such factors that might have a *potential* influence upon y. But Nature may limit the number of factors that have a nonnegligible *factual* influence to a relatively small number. (Haavelmo 1944, p. 24)

Thus, the relationship $y = F(x_1, \ldots, x_n)$ (see Eq. 4.13) explains the actual observed values of y, provided the factual influence of all the unspecified factors together were very small as compared with the factual influence of the specified factors $x_1, \ldots, x_n$.

> This might be the case even if (1) the unspecified factors varied considerably, provided their potential influence was very small, or if (2) the potential influences of the unspecified factors were considerable, but at the same time these factors did not change much, or did so only very seldom as compared with the specified factors. (Haavelmo 1944, p. 25)

However, "our greatest difficulty in economic research" does not lie in establishing simple relations, but rather in the fact the empirically found relations, derived from observation over certain time intervals, are "still simpler than we expect them to be from theory, so that we are thereby led to *throw away* elements of a theory that would be sufficient to explain apparent 'breaks in structure' later" (p. 26). This was the so-called problem of autonomy of economic relations. Some of these relations have very little autonomy because their existence depends upon the simultaneous fulfillment of a great many other relations. Highly autonomous relations were those that "describe the functioning of some parts of the mechanism *irrespective* of what happens in some *other* parts" (p. 28). This was the "principal

[6] In Boumans (2005), it is shown that Haavelmo's definitions of potential and factual influences can be represented in this way.

task of economic theory": to establish those relations that might be expected to possess as high a degree of autonomy as possible.

The problem is that it is not possible to identify the reason for the factual influence of a factor, say x_{n+1}, being negligible, $F_{n+1}\Delta x_{n+1} \approx 0$. We cannot distinguish whether its potential influence is very small, $F_{n+1} \approx 0$, or whether the factual variation of this factor over the period under consideration was too small, $\Delta x_{n+1} \approx 0$. We would like only to "throw away" factors whose influence was not observed because their potential influence was negligible to start with. At the same time, we want to retain factors whose influence was not observed because they varied so little that their potential influence was veiled.

The variation of x_{n+1} is determined by other relationships within the system. In some cases a virtually dormant factor may become active because of changes in the economic structure elsewhere. However, deciding whether a factor should be accounted for in the relationship under investigation should not depend on such changes. The relationship should be autonomous with respect to structural changes elsewhere.

How autonomous an equation is depends on our knowledge of the potential influence of each factor, F_i, which will inform us about the formal properties of the function F (see Haavelmo 1944, p. 24). Both Frisch and Haavelmo were pessimistic about whether it was possible to acquire knowledge about the autonomy of an equation through passive observation alone (therefore, they both advocated experiments in economics). However, the problem is not insurmountable if we use our knowledge about the time shapes of the phenomenon we want to explain. Facts about the time shape of y can be fed back to the form F of the relation being investigated. However, this only works if facts about the time shapes of the phenomenon are invariant and stable, and not temporary characteristics. As we discussed earlier, Haavelmo considered this strategy to be "very dangerous" and therefore abandoned it. Frisch also, mistakenly, believed that the time shapes were not sufficient to gain knowledge about the complete list of causal factors.

Unlike Haavelmo, Frisch and Tinbergen used (stylized) facts about the time shape of the business-cycle phenomenon. For Tinbergen, whether integral terms (hidden or not in the observations) should be included in the business-cycle mechanism depended on assumptions about the periodicity and amplitude of the business cycle. Frisch's Propagation and Impulse model (1933b) also used time shapes to gain knowledge about the business cycle's generating mechanism. However, his 1938 memorandum shows that the possibilities to identify the full list of causal factors depend on the connection between assumptions about the mathematical representation of the business cycle and assumptions about the mathematical representation of the explaining mechanism.

Haavelmo's design rules for doing econometrics were considered as an alternative to the experimental method of science (Morgan 1990, p. 262). However, although researchers at the Cowles Commission adopted Haavelmo's "blueprint" for econometrics (Morgan 1990, p. 251), they scrapped the term "autonomy" because it was believed that structural relations were autonomous (see Aldrich 1989). Only Haavelmo kept the term "autonomy" alive. If one looks at the index of

the 1950 monograph of the Cowles Commission, one finds only one reference to "autonomous relation," namely, to Haavelmo's chapter "Remarks on Frisch's Confluence Analysis and Its Use in Econometrics." In it he described an autonomous relation as a relation that "would hold regardless of whether or not other economic relations were fulfilled" (Haavelmo 1950, p. 263). This is again the case in the follow-up of the 1950 monograph, namely, the 1953 monograph *Studies in Econometric Method*.[7] There is only one reference of "autonomous equations" in the index, in Girshick and Haavelmo's chapter "Statistical Analysis of the Demand for Food: Examples of Simultaneous Estimation of Structural Equations." But now "structural" and "autonomous" seem to have converged.[8]

> Why is it that we are interested in one particular member of this infinite set of true systems? It is because, in setting up the original model, we believe that there is one particular system of equations that is a system of *autonomous*, or *structural* equations, that is, equations such that it is possible that the parameters in any one of the equations could *in fact* change, e.g., by the introduction of some new economic policy, *without* any change taking place in any of the parameters of the other equations. (Girshick and Haavelmo 1953, p. 106)

The reason why researchers at the Cowles Commission believed that the structural equations were autonomous is that the empirically found relationships may be simpler than theory would suggest. This could lead researchers to overlook potential influences. Moreover, there may be factors that were not only overlooked because they were not revealed empirically but were also not yet accounted for in theory. However, as passive observers "we cannot clear the data of such 'other influences', we have to try to introduce these influences in the theory, in order to bring about more agreement between theory and facts" (Haavelmo 1944, p. 19). Thus, it was assumed that the problem of autonomy could be avoided by building models to be as comprehensive as possible.

5 Conclusions

Originally, in business-cycle analysis, whether a potential causal factor was added to the business-cycle mechanism depended on whether it was theoretically as well as statistically and mathematically significant. Mathematical significance of a causal factor depended on considerations of whether the model containing that factor generated the appropriate facts about the phenomenon. Because feedback from the phenomena was cut off, the Cowles Commission approach was not to discover or test, but only to identify and measure.

---

[7] Both monographs are considered as containing the main body of the Cowles Commission's theoretical results (see Christ 1994, p. 32).

[8] For a more sophisticated account of this convergence, see Chao (2009), where he distinguishes between the invariance view and the theory view. Autonomy is equivalent to invariance but "does not say anything about the constitution of a system" (p. 71). So, the convergence is best described as the "invariance view of structure."

Hoover (1994) shows that the Cowles Commission approach, which he labels as "strong apriorism," is just one of the two strategies to secure invariance. In his view, it is better to see econometrics as an observational science such as astronomy. Because the observations made by econometrics instruments are observations of confluent relations, one should adopt the strategy of "weak apriorism." Theory guides observations but observation can suggest which elements of a theory are unsatisfactory: "Measurement requires prior theory; equally, theory requires prior measurement" (p. 73). Nevertheless, as Hoover emphasizes,

> econometric observations would be practically useless if they were completely unstable. We must, therefore, count on finding some stability and on supplementing econometric observations with other information, say institutional facts, if we are to distinguish between real changes in structure and our inability to focus our observations. (Hoover 1994, pp. 75–76)

It was this class of stable facts about business cycles that originally were used to solve the problem of autonomy.

References

Aftalion, Albert. 1927. The theory of economic cycles based on the capitalistic technique of production. *Review of Economic Statistics* 9(4): 165–170.

Aldrich, John. 1989. Autonomy. In *History and methodology of econometrics*, ed. N. de Marchi and C. Gilbert, 15–34. Oxford: Clarendon.

Aldrich, John. 1994. Haavelmo's identification theory. *Econometric Theory* 10: 198–219.

Boumans, Marcel. 1995. Frisch on testing of business cycle theories. *Journal of Econometrics* 67 (1): 129–147.

Boumans, Marcel. 1999. Built-in justification. In *Models as mediators*, ed. M.S. Morgan and M. Morrison, 66–96. Cambridge: Cambridge University Press.

Boumans, Marcel. 2005. *How economists model the world into numbers*. London/New York: Routledge.

Bullock, Charles J., Warren M. Persons, and William L. Crum. 1927. The construction and interpretation of the Harvard index of business conditions. *Review of Economic Statistics* 9(2): 74–92.

Chao, Hsiang-Ke. 2009. *Representation and structure in economics: The methodology of econometric models of the consumption function*. London/New York: Routledge.

Christ, Carl F. 1994. The Cowles Commission's contributions to econometrics at Chicago, 1939–1955. *Journal of Economic Literature* 32(1): 30–59.

Frisch, Ragnar. 1933a. Editorial. *Econometrica* 1: 1–4.

Frisch, Ragnar. 1933b. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honour of Gustav Cassel*, 171–205. London: Allen & Unwin.

Frisch, Ragnar. 1948. Repercussion studies at Oslo. *The American Economic Review* 38(3): 367–372.

Frisch, Ragnar. 1995. Statistical versus theoretical relations in economic macrodynamics. In *The foundations of econometric analysis*, ed. D. Hendry and M.S. Morgan, 407–419. Cambridge: Cambridge University Press.

Girshick, M.A., and Trygve Haavelmo. 1953. Statistical analysis of the demand for food: Examples of simultaneous estimation of structural equations. In *Studies in econometric method*, Cowles commission monograph 14, ed. W.C. Hood and T.C. Koopmans, 92–111. New York: Wiley. First published in 1947 in *Econometrica* 15.2: 79–110.

Haavelmo, Trygve. 1940. The inadequacy of testing dynamic theory by comparing theoretical solutions and observed cycles. *Econometrica* 8: 312–321.

Haavelmo, Trygve. 1944. The probability approach in econometrics. Supplement to *Econometrica* 12: 1–115.

Haavelmo, Trygve. 1950. Remarks on Frisch's confluence analysis and its use in econometrics. In *Statistical inference in dynamic economic models*, Cowles commission monograph 10, ed. T.C. Koopmans, 258–265. New York: Wiley.

Hanau, Arthur. 1928. *Die Prognose der Schweinepreise*, Vierteljahrshefte zur Konjunkturforschung, Sonderheft Berlin: Institut für Konjunkturforschung 7.

Hendry, David F., and Mary S. Morgan (eds.). 1995. *The foundations of econometric analysis*. Cambridge: Cambridge University Press.

Hoover, Kevin D. 1994. Econometrics as observation: The Lucas critique and the nature of econometric inference. *Journal of Economic Methodology* 1(1): 65–80.

Kalecki, Michal. 1935. A macrodynamic theory of business cycles. *Econometrica* 3: 327–344.

Karsten, Karl G. 1924. The theory of quadrature in economics. *Journal of the American Statistical Association* 19: 14–27.

Karsten, K.G. 1926. The Harvard business indexes – A new interpretation. *Journal of the American Statistical Association* 21: 399–418.

Keynes, John Maynard. 1939. Professor Tinbergen's method. *The Economic Journal* 49(195): 558–568.

Koopmans, Tjalling C., Herman Rubin, and Roy B. Leipnik. 1950. Measuring the equation systems of dynamic economics. In *Statistical inference in dynamic economic models*, Cowles commission monograph 10, ed. T.C. Koopmans, 53–237. New York: Wiley.

Magnus, Jan R., and Mary S. Morgan. 1987. The ET interview: Professor J. Tinbergen. *Econometric Theory* 3: 117–142.

Marschak, Jacob. 1934. The meeting of the econometric society in Leyden, September–October, 1933. *Econometrica* 2: 187–203.

Morgan, Mary S. 1990. *The history of econometric ideas*. Cambridge: Cambridge University Press.

Tinbergen, Jan. 1927. Over de Mathematies-Statistiese Methoden voor Konjunktuuronderzoek. *De Economist* 11: 711–723.

Tinbergen, Jan. 1928. Opmerkingen over Ruilteorie. *Socialistische Gids* 13(5): 431–445, 13.6: 539–548.

Tinbergen, Jan. 1931. Ein Schiffbauzyklus? *Weltwirtschaftliches Archiv* 34(2): 152–164.

Tinbergen, Jan. 1933a. L'Utilisation des Équations Fonctionnelles et des Nombres Complexes Dans les Recherches Économiques. *Econometrica* 1: 36–51.

Tinbergen, Jan. 1933b. *Statistiek en Wiskunde in Dienst van het Konjunktuuronderzoek*. Amsterdam: Arbeiderspers.

Tinbergen, Jan. 1935. Annual survey: Suggestions on quantitative business cycle theory. *Econometrica* 3(3): 241–308.

Tinbergen, Jan. 1939a. *Statistical testing of business-cycle theories I; a method and its application to investment activity*. Geneva: League of Nations.

Tinbergen, Jan. 1939b. *Statistical testing of business-cycle theories II; business cycles in the United States of America*. Geneva: League of Nations.

Chapter 5
Models of Mechanisms: The Case of the Replicator Dynamics

Till Grüne-Yanoff

Abstract The general replicator dynamics (RD) is a formal equation that is used in biology to represent biological mechanisms and in the social sciences to represent social mechanisms. For either of these purposes, I show that substantial idealisations have to be made – idealisations that differ for the respective disciplines. These create a considerable *idealisation gap* between the biologically interpreted RD and the learning interpretations of the RD. I therefore argue that these interpretations represent different mechanisms, even though they are interpretations of the same formal RD equation. Furthermore, I argue that this idealisation gap between the biological and economic models is too wide for the respective mechanisms to share a common abstract causal structure that could be represented by the general RD model.

1 Introduction

It has become fashionable in recent philosophy of science to explicate the use of scientific models by claiming that they represent mechanisms. In this chapter, I discuss the replicator dynamics (RD), an important model in biology and economics, and argue that it does not represent a mechanism. The argument proceeds in two steps. First, I show that even though the same RD model is employed in biology and economics, the different interpretations in these disciplines make it represent different mechanisms. Second, I argue that these different mechanisms do not instantiate a common, more abstract, mechanism. Rather, different kinds of idealisations are imposed on the RD model, depending on whether it is interpreted in economics or in biology. This opens an 'idealisation gap' between the different

T. Grüne-Yanoff (✉)
Avdelningen för Filosofi, Royal Institute of Technology (KTH),
Teknikringen 78 B, 100 44 Stockholm, Sweden
e-mail: gryne@kth.se

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_5,

biological and economic models, too wide for the respective mechanisms to share a common abstract causal structure that could be represented by the general RD model.

The chapter is structured as follows. Section 2 introduces the needed distinctions between mechanism sketches, abstract models and complete models on the one hand and particular mechanisms and abstract mechanisms on the other. Section 3 surveys the formal RD model and its derivation from evolutionary game theory. Section 4 discusses its use by population biologists, who intended it as a representation of biological mechanisms. In Sect. 5, I discuss economists' use of the same RD equation to represent social mechanisms and argue that these social mechanisms are distinct from the biological ones. Section 6 contains the main argument, showing that the biological and economic models are separated by an 'idealisation gap' too wide for the respective mechanisms to share a common abstract causal structure that could be represented by the general RD model. Section 7 concludes.

2 Models and Mechanisms

The notion of mechanism has had significant impacts on the way philosophers of science account for the use of models in the sciences, in particular in biology, economics and the neurosciences. According to these accounts, models explain because they represent the mechanism that produces the phenomenon to be explained (Craver 2006, p. 367). Models help in controlling the real world, because their mechanism representations enable modellers to answer counterfactual questions (Woodward 2002, p. S371). Finally, we can make true claims with models, because they correctly represent an isolated mechanism, even when they idealise the influence of many background condition (Mäki 2009, p. 30).

In each of these functions, models *represent* mechanisms. Whatever the specific definition of mechanism is (I will remain noncommittal here, as different incompatible definitions are extant and the detail of these does not matter for my purposes here), it is clear that mechanism is considered a part of the real world, characterised, for example, as 'material structures' (Craver and Kaiser 2013, p. 130) or a 'portion of the causal structure of the world' (Craver and Kaiser 2013, p. 141).

A mechanistic model may be designed to represent more or less details of a mechanism. Here authors have distinguished between mechanisms sketches, schemata and complete mechanistic models. A *sketch* is an 'incomplete model of a mechanism' (Craver 2006, p. 360). While characterising some parts, activities and features of the mechanism's organisation, it leaves blanks. These blanks are not necessarily visible, as they may be camouflaged by 'filler terms': terms like 'activate', 'inhibit' or 'produce' that indicate activity in a mechanism without detailing how the activity is carried out. Thus, there is more to the represented mechanism than a representing model sketch says.

On the other extreme, we have an *ideally complete* model. 'Such models include all of the entities, properties, activities, and organizational features that are relevant to every aspect of the phenomenon to be explained' (Craver 2006, p. 360). Even if completeness is relativised with respect to explanatory purpose, few, if any, such complete models can be found. More relevant, thus, seems the notion of a mechanism schema, which is a somewhat complete, but less than ideally complete, mechanistic model.

For a given mechanism, a mechanism sketch thus represents less of its features than a mechanism schema does. The sketch does so either by not at all specifying some features that the schema specifies (this is easier with formal models: a set-theoretic model, say, may stay silent about the colours of the objects it represents; a computer model may successfully evade specifying the weight of the structure it represents). Alternatively, sketches often specify certain features, but users of the sketch might exclude these features from the representational function of the sketch. That is, they declare these certain features to be *idealisations*. Scale models, for example, have many features, like size and weight and materiality, that are usually considered idealisations and hence not representations of the target object's properties. By either way, a mechanistic sketch represents less features of a given mechanism than a mechanism scheme does. Consequently, mechanism sketches are more *abstract* than mechanism schemata.

Abstraction is often thought of in relation to generality.[1] A mechanism sketch, then, is more abstract than a mechanism schema, because those properties described in the sketch are a proper subset of those described in the schema. Different mechanisms, described by different schemata, may therefore be described by one and the same sketch.

Such a view of mechanistic models is particularly plausible when seen from an 'exemplar' account of mechanisms. Such an account points out that mechanistic models often represent a particular, exemplary mechanism (Bechtel and Abrahamsen 2005, p. 438). Such exemplars or prototypes are particular tokens of causal structure in the world. A mechanistic model close to being ideally complete might represent just a single such exemplar. A mechanism sketch, on the other hand, might represent a large set of such exemplars. With increasing abstraction, mechanistic models get more and more general.

Scientists use exemplars and prototypes, according to Bechtel and Abrahamsen, in order to accommodate the subtle variations between related mechanisms. For example, they model a mechanism in wild-type *Drosophila* and then extrapolate from this prototype to mechanisms in other strains and species, all the while acknowledging that these are not identical mechanisms. In this view, explaining with a mechanistic model typically commences by explaining the phenomenon with

[1] Take, for example, Nancy Cartwright's Aristotelian account of abstraction: '*A* is a more abstract object than *B* if the essential properties, those in the description of *A*, are the proper subset of the essential properties of *B*' (Cartwright 1989, p. 214).

a prototype mechanism, which is then judged to be sufficiently similar to the mechanism that actually produced the phenomenon.

Yet this is not the only way how one can conceive of inferences between mechanisms. Instead of restricting one's ontology to concrete mechanisms that are only instantiated in one kind of organism, or one kind of social institution, one might also accept that there are *abstract mechanisms* that have many concrete instantiations in different kinds of organisms or institutions. This idea has been floated by some writers, who propose a sort of hierarchy of mechanisms. It is worthwhile quoting one such argument at length.

> Processes identified in the causal reconstruction of a particular case or a class of macro-phenomena can be formulated as statements of mechanisms if their basic causal structure (e.g., a specific category of positive feedback) can also be found in other (classes of) cases. The mobilization process observed in a fund-raising campaign for a specific project can, for instance, be generalized to cover other outcomes such as collective protest or a patriotic movement inducing young men massively to enlist in a war. A particular case of technological innovation like the QWERTY keyboard may similarly be recognized as a case in which an innovation that has initially gained a small competitive advantage crowds out technological alternatives in the long run. This is already a mechanism of a certain generality, but it may be generalized further to the mechanism of "increasing returns," which does not only apply to technological innovations but has also been used in the analysis of institutional stability and change … "Increasing returns," of course, is a subcategory of positive feedback, an even more general mechanism that also operates in the bankruptcy of a firm caused by the erosion of trust or in the escalation of violence in clashes between police and demonstrators. (Mayntz 2004, p. 254)

Central to this idea is that more abstract mechanisms exist in the same way as concrete mechanisms are said to exist. Abstract mechanisms are *instantiated* in more concrete mechanisms: Mayntz' positive feedback mechanism is instantiated in escalation of violence between police and demonstrators, in trust-erosion mechanisms and in increasing returns mechanisms. Mechanisms of different degrees of abstraction are also *nested*: positive feedback is instantiated in increasing returns, which in turn is instantiated in technological crowding out, which in turn is instantiated in the specific process that led to the dominance of the QWERTY keyboard.

According to this view, inferences between mechanisms do not go from prototypes to similar particular mechanisms, but they go through abstract mechanisms in the form of shared 'basic causal structure'. Explaining with a mechanistic model commences by explaining the phenomenon with an abstract mechanism and then showing that the mechanism that actually produced the phenomenon is an instantiation of the abstract mechanism.

Allowing for abstract mechanisms produces an ontological mirror image to the abstraction hierarchy of models. Unlike the exemplar account, which casts all models as more or less abstract representations of particular mechanisms, the abstract mechanism account allows models to represent both abstract and particular mechanisms. Consequently, what appears at first sight to be a mechanism sketch might either be an abstract representation of particular mechanisms or a nearly complete model of an abstract mechanism.

Prima facie, the abstract mechanisms account fits well with observed scientific practice. Scientists often speak about abstract causal structures as if they were real. They see patterns, structures and processes instantiated in various events that produce phenomena: for example, natural selection in the genesis of traits of many different organisms or positive feedback loops yielding dominance of certain set-ups in many institutions. They model these abstract patterns, structures and processes and suggest that these models represent something real.

Conversely, scientists sometimes question the legitimacy of abstract mechanistic models by arguing that an abstract mechanistic model is a mere sketch and not a representation of an abstract mechanism. For example, a paper glider might be a useful mechanism sketch of flight mechanisms in both birds and flying machines. But as parents will explain to their little paper pilots, this does not mean that bird flight and machine flight share the same basic causal structure. Rather, birds combine the function of providing both lift and thrust in their wings, while airplanes separate these functions. Such an explanation implicitly distinguishes between *genuine abstract models* that represent abstract mechanisms and *spurious abstract models* that are mere sketches of concrete mechanisms, to be filled in different and differentiating ways.

This is the problem that evolutionary game theorists face, too: they operate – amongst other formalisms – with the RD model. This model is very abstract: it is used to represent concrete mechanisms that clearly differ in some of their properties. Crucially for my question, the RD model is used to represent mechanisms both in economics and biology. The question thus arises whether the RD model represents the same abstract mechanism in both disciplines or whether it is a mere mechanism sketch that represents a set of disparate concrete mechanisms.

I argue that the RD is a spurious abstract model: a mere mechanism sketch that requires filling in to represent the relevant features of the respective biological and social mechanisms. As I will argue in Sect. 6, this 'filling in' of the RD follows discipline-specific paths that increase the idealisation gap between biological and social RD models. But before I can make that argument, I need to investigate the modelling projects in the two disciplines in more detail.

3 The Replicator Dynamics

Evolutionary game theory (EGT) investigates the compositional stability of a population as the result of interaction amongst its members. One of its most prominent modelling approaches derives a differential equation for the population composition from the game matrices that detail payoffs from interaction for each individual in the population. Thus, in contrast to classical game theory, EGT focuses not on decisions of individual players, but on properties of the whole population and on the effect of properties of previous populations on future population. This effect is represented through various population dynamics, first and foremost the replicator dynamics (RD).

Fig. 5.1 The general RD model

Let me describe the RD model in more detail. A *population* is a set of individuals. Individuals are programmed to play one strategy. A *strategy* is a complete plan of action for whatever situation might arise; this fully determines the player's behaviour. A *population state* is defined as the vector $x(t) = (x_1(t),\ldots, x_k(t))$, where each component $x_i(t)$ is the frequency of strategy i in the population at time t.[2] The replicator dynamics is a function that maps a population state at time t onto a population state at $t + 1$. It exists both as a discrete version, in which $x(t + 1) = f(x(t))$, and as a continuous version, in which for each i, $\mathrm{d}x_i/\mathrm{d}t = f(x(t))$.

The RD function relates to the interaction of individuals in the population through the following five steps. First, a population of individuals is presented and the variation of strategies in the population described in the population state. Second, in each period, every individual is paired at random with another individual from the population. These individuals play the strategies that they are programmed to play against each other. Third, a *game* is specified that members of the population play between each other. Commonly, this game is a two-player simultaneous-move game that for each player includes all strategies present in the population state. For each *strategy profile* (i, j) – a combination of strategy i of one player and strategy j of another player – the game specifies a *payoff* $u_k(i, j)$ for each player $k = \{1, 2\}$. Fourth, the payoff individual received from the interaction is interpreted as affecting the replication of this individual: how many individuals will play strategy i in the next period is proportional to how well individuals playing i in this period did vis-à-vis other individuals. Fifth, proportionality of replication and payoffs leads to differential representation of strategies in the population in the next period. Over many periods, this differential representation may lead to the convergence of stable state, in which differential representation of traits becomes stable over time, unless disturbed exogenously. Alternatively, differential representation might change in a regular fashion, for example, in regular oscillations or circles. Tracking the outcome of the dynamics over time reveals such stability or regularity results. Figure 5.1 depicts these five steps graphically.[3]

Mathematically, these steps are represented as follows. Given a population state $x(t)$, the expected payoff to any pure strategy i in a random match is $u(i, x)$: an

[2] The population state is formally identical to a mixed strategy. Its support is the set of strategies played by individuals in the population.

[3] These and the following graphs are schematic representations of models – of the formal RD equation *and* its respective interpretations. I use these graphs in order to make comparison between the different models more palpable.

individual that plays i against a randomly drawn opponent faces every strategy present in the population with the associated frequency with which that strategy occurs in the population. This is formally identical to this individual playing against an opponent who plays a mixed strategy $x(t)$. The associated population average payoff is $u(x, x) = \Sigma_i x_i * u(i, x)$.

The frequency of strategy i changes to the degree that the expected payoff $u(i, x)$ differs from the population average payoff $u(x, x)$. If $u(i, x)$ is greater than $u(x, x)$, the number of individuals playing i in the next period will grow more than the population average. If $u(i, x)$ is smaller than $u(x, x)$, the number of individuals playing i in the next period will shrink more than the population average. This relative growth is assumed to be linearly proportional to the difference between strategy payoff and the population average payoff.[4] Consequently, the continuous RD is specified as follows:

$$\frac{dx_i}{dt} = [u(i,x) - u(x,x)] * x_i \quad \text{(Weibull 1995, p. 72)} \tag{5.1}$$

That is, the change in x_i's population share is determined by x_i's current population share and the difference between its expected payoff and the population average payoff.

Through analysis of a phase diagram of these dynamics, convergent trajectories, stable states and regular changes can be identified. Under the biological interpretation, regular changes identify the temporal predominance of certain traits in the population, while stable states identify results of adaptation of organisms to their environment.

4 The Biological RD

The RD was first derived in the late 1970s and quickly became the most prominent model of evolutionary dynamics in EGT.[5] The RD is derived from EGT by implicitly presupposing EGT to describe an underlying biological mechanism. The core idea of EGT in biology is that organisms often find themselves in strategic situations, in which the fitness-relevant outcome of their behaviour at a certain time depends on the behaviour of the other organisms in the population at that time. The fitness of an organism thus is influenced by the frequency of behaviour in that population. Consequently, there is a systematic relationship between the kind of

[4] The relation between proliferation and payoffs characterises different classes of selection dynamics. While a linear relation characterises the RD, wider classes are characterised by payoff positivity and payoff monotonicity, respectively (Weibull 1995, pp. 139–152). Yet the RD, which takes payoffs to represent fitness differences, is the most prominent selection dynamic in EGT and therefore will be discussed here.

[5] For a historical survey, see Grüne-Yanoff (2011a).

Fig. 5.2 The biological interpretation of the RD

composition of a population at a certain time and the differential reproduction of the respective strategies in that population at the next time step.

In particular, the biological interpretation gives causal substance to the formal five steps of the RD model above. First, individuals are interpreted as organisms and their strategies as certain inheritable behavioural traits. Second, organisms interact, for example, by fighting, mating, exchanging or collaborating. In this interaction, each organism exhibits the behavioural trait it is endowed with. Third, each organism receives an outcome from that interaction – for example, territory, food and mating partner – depending on its own behavioural trait and that of the organism it interacted with. Fourth, this outcome determines the number of offspring the organism has in the next period. Fifth, weighing growth of each trait by overall population growth yields the differential reproduction of each kind of organism.

The RD model is used to represent this mechanism. But it is not the formal RD model alone that performs this representational function, but rather a biologically interpreted RD model. In particular, the causally relevant properties are not found in the mathematical expressions of RD, but in its biological interpretation. This interpretation has turned the RD formalism into representations of specific causal forces and specific arrangement of these forces. Figuratively speaking, it fills in the black boxes of Fig. 5.1 to yield a causal process from population at t to population at $t + 1$, through interaction and differential reproduction, as shown in Fig. 5.2.

The biological interpretation of the mathematical expressions specifies the causal properties that bring about the result and that tell us *how* the population state changes from $x(t)$ to $x(t + 1)$. Hence, the mathematical RD model *in conjunction with* the biological interpretation represents the causal process from initial conditions to specific outcome, not the formal model alone.[6] The RD model is thus a mere sketch of the biological mechanisms it represents, as various gaps are filled in by the biological interpretation. Let me therefore distinguish the – more sketchy – formal RD model from the – less sketchy – biologically interpreted BRD model.

Even if it is less sketchy than the RD, the BRD does not represent a particular mechanism. Instead, it is a schema that represents mechanisms differing in many

[6] I have elsewhere argued that models consist of a formal structure *and* a story (Grüne-Yanoff and Schweinzer 2008). In the case I am discussing, the RD equation (5.1) constitutes the formal structure. The biological interpretation of its terms, and the account of interaction yielding a fitness-relevant outcome, leading to differential reproduction, constitutes the story.

details. BRD *abstracts* from these details. For example, it deals with strategies, abstracting from any concrete content of behavioural plans. Furthermore, it deals with generic organisms, not specific species or individuals. Finally, it abstracts from any differences between organisms, as when it assumes that organisms have the same fitness base rate.

Besides omitting and hence abstracting from many features, BRD also makes many specific assumptions about the processes it purportedly represents, even though these assumptions are likely to be false of many of these processes. Typical *idealisations* of this sort include the assumption that organisms match and interact with others randomly, hence idealising possible local interactions and network structure. It also idealises inheritance, assuming away epigenetic effects and sexual reproduction. BRD thus is an abstract and idealised representation of a process that supposedly can be found in many different concrete instantiations.

That it is seen as a representation of one abstract mechanism lies in the success of its application: many phenomena – in particular those involving frequency-dependent selection, like sex ratios, fighting behaviour or cooperation – have been successfully explained by reference to this abstract mechanism.

5 The Social RD

From the 1980s onwards, social scientists have increasingly adopted EGT for their own explanatory purposes. In particular, EGT has been used in order to explain the evolution of social institutions, in particular of conventions, norms and fairness preferences.

Sometimes, social scientists not only employed the general RD model but also resorted to its biological interpretation. For various reasons, this is today not considered adequate for most social science purposes.[7] Instead, specifically social interpretations of the RD have been proposed. These social interpretations represent mechanisms that account for the social interaction between individuals and for the social replication of these individuals' traits. A particularly important class of such mechanisms has been described as *learning*. Learning is an extremely open concept, and in the following I will only concentrate on those kinds of learning that are

[7] These difficulties spring from many sources; I just want to sketch three reasons here. First, while animals largely exist on the subsistence level, humans mainly do not. It is consequently much less clear what the causal effect of, say, adherence to norms is for survival and reproduction in humans, than what the causal effect of daily competition for food, shelter and mating opportunities is for survival and reproduction in nonhuman animals. Secondly, while it may be plausible that some basic animal behaviour is encoded in ways that can be inherited through reproduction, it is much less clear that complex human behavioural characteristics, like compliance with norms, can. Thirdly, the speed of cultural evolution is often much higher than human reproduction. Conventions in small groups, for example, can emerge or change within days, thus making reference to player reproduction inadequate. For these as well as other reasons, strategy replication often has to be thought of in ways independent of player reproduction.

Fig. 5.3 The reinforcement interpretation of the RD

purportedly described by the RD. Within that class, three kinds of learning can be distinguished: reinforcement, imitation and belief learning.

In *reinforcement learning*, a player's received payoffs from past interactions are her only feedback information. That is, the probability of a strategy to be played in the future is proportional to the success it gave the player in the past. Börgers and Sarin (1997) present a well-known model of such learning, which conforms to the replicator dynamic. In their model, a player at stage n plays a mixed strategy $P(n) = (P_1(n),\ldots, P_J(n))$ that includes all possible pure strategies $S_1,\ldots, S_J$ in the population. The player i observes the (pure) strategy S_k and its payoff $u_i(S_k, S_{-k})$, normalized to lie between 0 and 1, that is realised when she implements her mixed strategy against other players playing S_{-k}. She then 'learns' by adjusting the weight P_k of S_k in her mixed strategy in proportion to the payoff that S_k gave her by the following rule:

$$P_k(n+1) = u_i(S_k, S_{-k}) + (1 - u_i(S_k, S_{-k})) * P_k(n) \tag{5.2}$$

$$P_{k'}(n+1) = (1 - u_i(S_k, S_{-k})) * P_{k'}(n) \quad \text{for all } k' \neq k$$

For the specific case of only two actions, the expected movement of action probabilities based on this model equals the RD, rescaled by a constant (Börgers and Sarin 1997; Börgers et al. 2004). More generally, if the decision-maker uses Cross' learning rule, (and satisfies the model's other requirements), then the learning dynamics satisfies *monotonicity* and *absolute expediency* (Börgers et al. 2004). Both of these properties are also satisfied by the RD. Thus, there is an analogy between Cross learning and the RD. Börgers et al. (2004, p. 358) conclude from this that their results 'strengthen the case of the use of RD dynamics in contexts where learning is important'. They also speculate that it may be possible to adopt their results 'to an evolutionary setting' (Börgers et al. 2004, p. 400) but refrain from making any specific claims about this.

The reinforcement interpretation of the RD model can be graphically presented as shown in Fig. 5.3.

This interpretation differs in a number of features from BRD. It commences with agents playing mixed strategies (where all organisms share the same support) rather than pure strategies. These strategies are not inherited, but adopted and adjusted by the agents. It does not interpret payoffs as fitness, but as subjectively evaluated outcomes. It is these subjective evaluations that cause the agent's adjustment of her own strategies. And it is this adjustment, and not differential reproduction, that constitutes differential representation in the population.

In *imitation* learning, players occasionally sample other players in the population and learn about their strategy and the payoff they realised in the last round. They

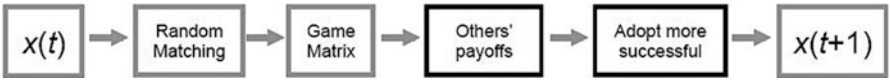


Fig. 5.4 The imitation interpretation of the RD

then switch their strategies according to the following rule: if in a population with state $x(t)$ the agent i's payoff is $u_i(x)$, and the agent samples an agent j with payoff $u_j(x)$, the agent switches with probability[8]

$$q_i = \max\{0, b(u_j(x) - u_i(x))\} \tag{5.3}$$

(Schlag 1998, p. 150, cf. also Weibull 1995, pp. 152–161). That is, she retains her strategy if her realised payoffs are greater than that of the sampled player. Otherwise, she adopts the strategy of the sampled player with a probability proportional to the difference between her and the sampled payoff. For this reason, such models are sometimes seen as closely related to the meme concept (Börgers 1996). The resulting population dynamic – in a large but finite population – is approximated by a deterministic dynamic that is analogous to the discrete RD (Schlag 1998, p. 152). Schlag furthermore points out that his model arrives at this result solely based on individual information and induced performance, while reinforcement learning models discussed above 'contain axioms concerning the functional form of a desirable learning curve' (Schlag 1998, p. 153).

The imitation interpretation of the RD model can be graphically presented as shown in Fig. 5.4.

This interpretation differs in a number of features from BRD. Although agents here also play pure strategies, it drops the heritability of strategies. Like the interpretation of Fig. 5.3, it does not interpret payoffs as fitness, but as subjectively evaluated outcomes. But unlike the reinforcement schema, the imitation schema models agents as evaluating not only their own but also others' outcomes. It is these subjective evaluations that may cause the agent to adopt another agent's strategy if she finds it more successful than her own. And it is this conditional adoption, and not differential reproduction, that constitutes differential representation in the population.

The previous two kinds of models cast learning as an influence of past payoffs (either of the player herself or of other players) on future behaviour. *Belief learning*, in contrast, models learning as experience influencing beliefs, and only through this influence, there is an indirect effect on behaviour. Hopkins (2002, p. 2144) has termed the particular kinds of belief learning modelled with EGT 'hypothetical reinforcement'. This is because players are modelled as calculating what they

[8] The function b ensures that the difference are normalised – that is, for any payoffs u_i, u_j in the population, $0 \leq b(u_j(x) - u_i(x)) \leq 1$.

would have received had they chosen some other action, on the basis of knowledge of their own payoff matrices and observations of their opponents actions.

This approach reinterprets EGT in general, and strategy selection in particular, as a theory of individual mental processes. Under this interpretation, all references to payoffs of others in a given environment are understood counterfactually as the payoffs that one would get in that environment if one adopted the other's strategy. For example, if a player knows the payoffs of each strategy profile, and knows the frequency with which strategies are played in the population, she can compare the expected payoffs of these strategies based solely on her own preferences. Having compared the strategies according to her own preferences, she can then choose that strategy that is either better than the current strategies or a best reply to her belief about the frequencies in the population. Variants of such models have been proposed by Sugden (1986), in Kandori et al.'s (1993) 'stochastic fictitious play' and in Young's (1993) 'adaptive play'.

Take, for example, Young's (1993) model. He defines play at time t as the strategy-tuple $s(t) = (s_i(t),\ldots, s_n(t))$, consisting of each player's strategy choice at time t. At period t, each player samples the past play h of a certain number of past periods. From this sample, the player constructs strategy-tuple s_h by weighing the past play in some way. Strategy-tuple s_h constitutes her estimate how other players will play in the next period. Thus, for the next period, agent i chooses s_i as the best reply to s_h. By choosing s_i, the player replaces the history of past play h with a new history h', in which the earliest period is removed and the most recent play added. This yields a process

$$P^0_{hh'} = \Pi p_i(s_i|h) \tag{5.4}$$

Where $P^0_{hh'}$ is the probability of moving from h to h′, determined as the product of the player's probabilities of choosing s_i given sample h. Young calls this process *adaptive play*.

Young's model is an example of what I call a *mental play* interpretation of EGT. What is relevant for a certain strategy to be selected no longer is the effect of actual interaction in a real population, but rather the consequence of an individual player evaluating various options, based on her subjective value criteria and her beliefs what her opponents will play. She forms these beliefs from her perception of and through reasoning about others' past play. She chooses her strategy by mentally representing her various options in the anticipated environment, figuring out the consequences of these counterfactual scenarios and choosing the one with the outcomes she values better or best.

Consequently, because the causal relation is between interaction and individuals' mental attitudes, no interpersonal payoff comparison is necessary. Players only observe their own payoffs from past play, and this affects only their own attitudes towards future play. Effects on aggregate properties are not directly modelled.

If noise is introduced into models of fictitious play, the expected motion of fictitious play becomes a form of noisy replicator dynamic (Hopkins 2002, p. 2149).

Fig. 5.5 The belief-learning interpretation of the RD

The only way that learning behaviour generated by stochastic fictitious play differs from the population dynamics of the two previous models is that they may differ in speed of passage along similar paths.

The belief-learning model can be graphically presented as shown in Fig. 5.5.

This interpretation differs in a number of features from BRD. It commences with agents playing mixed strategies (where all organisms share the same support) rather than pure strategies. These strategies are not inherited, but adopted and adjusted by the agents. Furthermore, it makes the crucial extra assumption that the whole population and all its strategies and payoffs are mentally represented by each organism. Based on this representation, the agent estimates how other players will play in the next period. Furthermore, the schema does not interpret payoffs as fitness, but as subjectively evaluated outcomes. Based on the estimation of others' future play, and her own subjective evaluations, the agent then chooses her action as a best reply. It is this deliberation, and not differential reproduction, that causes differential representation in the population.

6 Relating the Mechanisms

It should be clear from the comparison of the previous section that the three learning mechanisms are not identical with what the BRD represents. In particular, what kind of strategies individuals play, how payoffs are realised and what information and what mental capacities individuals employ in replicating strategies differ considerably between BRD and the learning interpretations of the RD (as well, to a lesser extent, between these interpretations themselves). Thus, the BRD and the respective learning interpretations of the RD represent *different* mechanisms, even though all these mechanisms are represented by the same RD model.

The RD model thus appears in the first instance as a highly abstract mechanism sketch. It is used to represent different kinds of mechanisms, but for each of these representation tasks, it needs to be filled in with a more domain-specific interpretation or story.

Nevertheless, one might still want to defend the claim that the general RD represents one mechanism – namely, by arguing that the BRD and the learning mechanisms all instantiate a more abstract mechanism and that this abstract mechanism is represented by the general RD model.

This idea seems prima facie plausible, particularly when one recalls that the BRD and the learning models themselves are abstract representations of mechanisms. As I discussed in Sect. 4, the BRD abstracts from any concrete content of behavioural plans, from specific species or individuals and from any differences between

Fig. 5.6 *Fliegende Blätter* (Oct. 23, 1892, p. 147, Nr. 2465)

organisms. For example, it omits representations of how organisms reproduce and instead describes the stage as a general process of reproduction in all its possible forms. So if the BRD is an abstract representation of a class of mechanisms, why should the general RD not be an even more abstract representation?

Furthermore, the fact that both the BRD and the learning models use the RD for their representational tasks seems to provide evidence that indeed there is an abstract mechanism instantiated both in the more concrete biological and social mechanisms and that this more abstract mechanism is represented by the RD model. Because the RD contains those features shared by the BRD and the learning models, it might seem plausible to conclude that the RD represents that abstract causal structure shared by the biological and the social mechanisms.

Against this appearance, I will now argue that the general RD model is *not* a representation of an abstract mechanism, instantiated by both the biological and the social mechanisms. Rather, the way the two disciplines 'fill in' the RD model in order to represent their respective mechanisms differs considerably. Users of the RD model, when filling it in, make systematically different kinds of idealisations, depending on whether it is interpreted in economics or in biology. This leaves little to be shared between the respective represented mechanisms – little that could be represented by a single RD, however interpreted. Instead, the RD model faces an *idealisation gap*: it can be interpreted *either* biologically *or* in one of the learning senses, but it cannot be interpreted to capture the essence of all, because there is little essence to capture. To clarify my argument, let me illustrate it with a joke.

The joke's not mine – it was published 120 years ago in the *Fliegende Blätter*, a German satirical weekly. Most philosophers know its subject, the duck-rabbit, from Wittgenstein's discussion of aspectual perception or from Kuhn's discussion of a paradigm shift. What those discussions ignore is the way the joke was posed, as shown in Fig. 5.6. The German headline reads 'which animals are most similar?', and the answer is 'rabbit and duck'.

The author of this little vignette thus did not solely intend to entertain with the *Gestalt* shift, but rather used this shift in order to infer an obviously absurd and hence satirical conclusion: because the same image represents both a rabbit and duck, it is suggested, we must conclude that rabbit and duck are indeed most similar.

Obviously, this inference is absurd for a number of reasons. I want to focus here on a rather subtle one, namely, that the same image relates to the two objects it supposedly represents in different ways. When we use the above image as a representation of a rabbit, we make certain kinds of idealisation. For example, we idealise the size of the rabbit's mouth and nose, as well as the shape of its ears. When we use the image as a representation of a duck, however, we make *different* idealisations: the back of a duck's head looks different, and it has different markings on its feathers. Thus, when using the image to represent either the one or the other, we make different allowances for which part of the image may not be representationally accurate. The ingenuity of the draughtsman lay in creating one image that allowed us to make the respective idealisations in such a way that it can function either as a representation of a rabbit *or* a duck. By making these different idealisations, we adapt the image for its respective uses. Although a duck shares some features with the image, and a rabit also share some features with the image, *these are not the same features*. Thus understood, there is little reason to believe in the similarity of rabbits and ducks because they are representable with the same image.

The same holds for the RD model. To use a model as a representation, we always have to make some idealising assumptions. But when interpreting the RD model biologically, we make idealisations that systematically differ from those we make when interpreting the RD model socially. Let me list some of these differences.

First, all three learning models require that players in some way identify actions and strategies – either of their own or possibly of others. If agents could not identify strategies in this way, they would not be able to link a diagnosis of 'success' with the choice of a successful strategy. This stands in contrast to the biological model, where the strategy notion only fulfils a theoretical role: differential reproduction does not require that the organism identify the strategies.

This additional requirement pushes these learning models beyond a simple notion of copying. Rather, it involves the ability to attribute goals and intentions. 'Something other than copying is taking place' (Sperber 2000, p. 171), and this other factor may have the power to lead the process in directions that mere copying would not. Yet such factors are idealised away in all of the three learning models.

Second, unlike the biological model, the learning models make specific assumptions about the learning rules players employ, at the exclusion of other, possible rules. In the biological model, if the payoffs are interpreted as fitness, there is a natural justification for a linear relationship between payoffs and differential reproduction. Yet in the learning models, specific imitation and reinforcement rules have to be chosen to arrive at a linear relationship.

> Other imitation rules – as plausible as Schlag's – yield processes different from any biological ones. (Börgers 1996, p. 1383)

This is even more obvious with respect to belief learning. For example, choice of different reasoning principles or heuristics may lead to different beliefs about strategies, strategy outcomes, etc., even when based on the same actual interactions. This sensitivity of the population dynamic to the specifics of the learning rules increases the 'idealisation gap' between the biological and the learning models.

Third, and related to the previous point, all learning models have to make strong assumptions about players not making mistakes – they never switch from a better to a worse strategy. This is a real possibility in all learning models – as agents have to actively identify strategies, associate payoffs with them and choose their own actions on that basis – while it has no significance in the biological model. The way this is dealt with usually involves taking expected values. Averaging this way over the possible behaviours of an agent idealises the influence of players' mistakes away: even if there is a positive probability that a player will switch from better to worse, on average the player will not (cf. Gintis 2000, p. 192).

Fourth, stochastic fictitious play models face the particular problem of excessive time horizons. As Sobel starkly puts it,

> the long-run predictions [of stochastic fictitious play] only are relevant for cockroaches, as all other life forms will have long been extinct before the system reaches its limits. (Sobel 2000, p. 253)

To turn the stochastic belief-learning models into representations of social mechanisms, the time horizons thus must be idealised.

Fifth, the imitation learning model faces the particular problem of requiring interpersonal comparisons of utility (Grüne-Yanoff 2011b). The biological RD model does that, too – yet while this requirement is innocuous under the fitness interpretation, it is highly problematic when payoffs are interpreted as numerical representations of preferences. Thus, this extra requirement constitutes an important difference between the belief-learning models and the other models discussed here.

Certain substantial idealisations need to be taken also when the RD model is interpreted biologically. A different set of substantial idealisations needs to be taken when the RD model is interpreted socially. By making these different idealisations, we adapt the model for its respective representative uses. This is standard scientific practice: most, and possibly all, model uses involve idealisations.

Yet when the same formal structure is employed to construct different, more specific mechanistic models, and each of these models involves different idealisations, one has to be careful when inferring purported similarities between these different mechanisms based on the common formal structure. Like the duck-rabbit, the RD equation is adapted for its respective representative tasks. In the course of each adaptation, certain features of the RD are drawn on – others are accepted as useful or at least harmless idealisations. Which features are drawn on and which are accepted as idealisations differ with each adaptation. The mechanism that each adaptation of the RD represents is substantially different from each other and does not share any or little causal structure between each other. Thus, there is

no abstract mechanism that is instantiated by the biological and learning mechanisms, and consequently the RD cannot represent such a mechanism.

7 Conclusions

The general RD is a model that is used in biology to represent biological mechanisms and in the social sciences to represent social mechanisms. Substantial idealisations have to be made for these purposes – idealisations that differ for the respective disciplines. These create a considerable idealisation gap between the BRD and the learning interpretations of the RD. This gap is sufficiently large to conclude that the general RD does not represent an abstract mechanism that subsumes both the biological and the social cases. Just like the duck-rabbit image does not represent the essence of both duck and rabbit, but rather either a duck or a rabbit (depending on what idealisations one accepts), so the general RD represents either biological or social mechanisms, but not the shared causal structure of both.

References

Bechtel, W., and A. Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Börgers, T. 1996. On the relevance of evolution and learning to economic theory. *The Economic Journal* 106: 1274–1385.

Börgers, T., and R. Sarin. 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* 77: 1–14.

Börgers, T., A. Morales, and R. Sarin. 2004. Expedient and monotone learning rules. *Econometrica* 72: 383–405.

Cartwright, N. 1989. *Nature's capacities and their measurement*. Oxford: Clarendon.

Craver, C.F. 2006. What mechanistic models explain. *Synthese* 153: 355–376.

Craver, C.F., and M.I. Kaiser. 2013. Mechanisms and laws: Clarifying the debate. In *Mechanism and causality in biology and economics*, ed. Hsiang-Ke Chao, Szu-Ting Chen, and Roberta L. Millstein, 125–145, Dordrecht: Springer.

Gintis, H. 2000. *Game theory evolving*. Princeton: Princeton University Press.

Grüne-Yanoff, T. 2011a. Models as products of interdisciplinary exchange: Evidence from evolutionary game theory. *Studies in History and Philosophy of Science* 42: 386–397.

Grüne-Yanoff, T. 2011b. Evolutionary game theory between interpersonal comparisons and natural selection: A dilemma. *Biology and Philosophy* 26: 637–654.

Grüne-Yanoff, T., and P. Schweinzer. 2008. The roles of stories in applying game theory. *Journal of Economic Methodology* 15(2): 131–146.

Hopkins, E. 2002. Two competing models of how people learn in games. *Econometrica* 70: 2141–2166.

Kandori, M., G. Mailath, and R. Rob. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61: 29–56.

Mäki, U. 2009. MISSing the world: Models as isolations and crediblesurrogate systems. *Erkenntnis* 70(1): 29–43.

Mayntz, R. 2004. Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences* 34: 237–259.

Schlag, K. 1998. Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory* 78(1): 130–156.

Sobel, J. 2000. Economists' models of learning. *Journal of Economic Theory* 94: 241–261.

Sperber, D. 2000. An objection to the memetic approach to culture. In *Darwinizing culture: The status of memetics as a science*, ed. Robert Aunger, 163–174. Oxford: Oxford University Press.

Sugden, R. 1986. *The evolution of rights, cooperation, and welfare*. Oxford: Basil Blackwell.

Weibull, J.W. 1995. *Evolutionary game theory*. Cambridge, MA: M.I.T. Press.

Woodward, J. 2002. What is a mechanism? A counterfactual account. *Philosophy of Science* 69: S366–S377.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61(1): 57–84.

Chapter 6
Experimental Discovery, Data Models, and Mechanisms in Biology: An Example from Mendel's Work

Ruey-Lin Chen

Abstract The aim of this chapter is to argue that there are experimental discoveries that could have been made *independent of* theories. I will explore the questions of whether there are experimental discoveries and, if so, what counts as an experimental discovery and what the relation is between experimental discovery and the discovery of a mechanism. Gregor Mendel's work on peas will be taken as the main example. Frederick Griffith's experiment with *Pneumococcus* bacteria in mice and Hans Driesch's experiment on sea urchin embryos will be discussed as foils. I conclude that an experimental discovery can be identified and recognized by the following conditions: (1) An experimenter must propose *data models* to reveal significant phenomena, (2) *no* established theories can predict and explain the phenomena, and (3) the experimenter must envisage searching for *underlying mechanisms* for the phenomena, whether or not he or she proposes correct mechanistic explanations. I also argue that experimental discovery usually precedes and is a prerequisite for the discovery of mechanism. It plays a role in three ways: *organizing* data into significant phenomena, *producing* the need and motivation to discover mechanisms, and *constraining* the direction for construction of theoretical hypotheses.

1 Introduction

In the philosophy of science, the problem, research foci, and terms related to scientific discovery have evolved over time. In the beginning, philosophers of science had to resist the dominant idea in the field—that of the distinction between discovery and justification—and argue for the philosophical significance of

R.-L. Chen (✉)
Department of Philosophy, National Chung Cheng University,
168, University Rd, Min-Hsiung, Chia-Yi, Taiwan
e-mail: pyrlc@ccu.edu.tw

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_6,

discovery. Thomas Nickles (1980a) convincingly argued against the old dichotomy, showing that there were abundant philosophical problems related to discovery. Influenced by the historical approach in the 1970s, philosophers during the 1980s and 1990s explored the emergence of new theories in scientific changes (Nickles 1980b, c; Darden 1991). In addition, they searched for heuristic strategies leading to the generation of plausible hypotheses (Schaffner 1974, 1993; Nickles 1987; Darden 1991; Kleiner 1993; Bechtel and Richardson 1993). After the mid-1990s, a number of philosophers adopted the term "scientific reasoning" instead of "the logic of discovery," exploring patterns and strategies of reasoning in the discovery of theories (Darden 1991, 2006; Magnani et al. 1999; Bechtel 2006). In spite of the change of foci and terms, these works largely addressed the methodological question: How do scientists produce plausible hypotheses and reliable theories?

In the mid-1990s, some advocates of discovery shifted their interest to the metaphysical question: What exactly is discovered in a scientific discovery? The problem of the discovery of mechanisms was spotlighted. Philosophers proposed a new mechanistic philosophy, debated the conception of mechanism, investigated the relations between mechanism and other topics, and explored ways to discover mechanisms in biology (Bechtel and Richardson 1993; Glennan 1996, 2002, 2005; Machamer et al. 2000; Darden 2002, 2005; Craver 2002, 2005; Bechtel and Abrahamsen 2005; Bechtel 2006). These scholars are usually called the "new mechanists." It is easy to see the relevance of the metaphysical problem to methodology, because the goal of producing plausible hypotheses and reliable theories in many fields of science (probably excluding mathematical physics) is to discover mechanisms. Lindley Darden (2006) convincingly proposed an integrative methodology, connecting reasoning strategies with the discovery of mechanisms.

Despite the fruitful work that has taken place on scientific discovery, there are, in my view, still two deficiencies: First, the scope that philosophers have explored is so restricted that they have paid little attention to other sorts of scientific discovery. Second, philosophers seem to have left untouched the problem of the nature of scientific discovery.

Scientists and historians of science usually use the term "discovery" to refer to finding new phenomena, new entities, hidden structures, hidden patterns, and mechanisms. For instance, the discoveries of the photoelectrical effect, the electron, the helical structure of DNA, and the correspondence between specific amino acids and specific nucleic acid codons (the genetic code) are usually regarded as great events in the history of science. The new mechanists may think that the recognition of the discoveries of new phenomena, entities, and structures relies on discovering relevant underlying mechanisms. This view is plausible. However, we still need to carefully investigate the way that underlying mechanisms are used to recognize new discoveries.

Scientists always make discoveries by some means, for instance, theoretical predictions, model-based reasoning, experiments, or observations. The planet Neptune was discovered by theoretical computation and observation, the glow of the cathode ray by experimental instruments, the mechanism of protein synthesis by a set of experiments and theoretical models, and so on. Scientists have made a large

amount of discoveries by testing the predictions of a theory. For instance, the curved path along which light passes through a large gravitational field had been exactly predicted by the theory of relativity before it was discovered. The history of science seems to be filled with great discoveries made by confirming the predictions of overarching theories.

Moreover, a number of philosophers believe that even if an experimental or observational result is not predicted by any established theory, its recognition is based on theoretical interpretations. They would suppose that, for example, the discovery of oxygen was recognized only when Antoine Laurent Lavoisier developed a new theory of combustion and the elements. If all observational and experimental results require a theoretical basis, as theory-oriented philosophers believe, then people might conclude that no discoveries can be made without theories. Is this true? Are there no discoveries that could be made simply by conducting an experiment? Could discoveries be made without the engagement of theories?

To date, philosophers have focused on the role and contribution of theories when investigating various cases of scientific discovery. Recently, a few philosophers of science (Hacking 1983; Mayo 1996; Galison 1998; Waters 2004) have discussed experiments in their own right but have paid little attention to the way experiments lead to new discoveries. By contrast, philosophers who analyze discoveries have not neglected experiments but have focused mainly on the experimental testing of theoretical hypotheses rather than the experimental discovery of phenomena or entities (Schaffner 1993; Craver 2002; Darden 2006). As a result, the problem of whether there exist experimental discoveries has not yet been explored. Of course, theory-oriented philosophers may think that recognition of an experimental discovery relies on discovering an explanatory theory. Again, we need to carefully investigate whether this is so.

The aim of this chapter is to argue that there are experimental discoveries that could have been made *independent of* theories. I will explore the questions of whether there are experimental discoveries and, if so, what counts as an experimental discovery and what the relation is between experimental discovery and the discovery of a mechanism. Gregor Mendel's work on peas will be taken as the main example. Frederick Griffith's experiment with *Pneumococcus* bacteria in mice and Hans Driesch's experiment on sea urchin embryos will be discussed as foils.

My argument proceeds in the following order: First, I specify problems with the identification of scientific discoveries and the recognition of the discoverers and the discovered in Griffith's and Driesch's experiments. Second, I introduce historical controversies about Mendel's discovery and argue that what Mendel discovered is two data models extracted from his experiments with peas. Following this argument, I discuss the relationship between experimental discoveries and mechanisms.

I conclude that an experimental discovery can be identified and recognized by the following conditions: (1) An experimenter must propose *data models* to reveal significant phenomena, (2) *no* established theories can predict and explain the phenomena, and (3) the experimenter must envisage searching for *underlying mechanisms* for the phenomena, whether or not he or she proposes correct

mechanistic explanations. I also argue that experimental discovery usually precedes and is a prerequisite for the discovery of mechanism. It plays a role in three ways: *organizing* data into significant phenomena, *producing* the need and motivation to discover mechanisms, and *constraining* the direction for construction of theoretical hypotheses.

2 Problems with the Identification and Recognition of Experimental Discoveries

In 1928, Frederick Griffith conducted a novel experiment on two strains of *Pneumococcus* bacteria, known as the smooth (S) strain and the rough (R) strain for the formation of colonies with smooth or rough surfaces, respectively. It was known that the S-strain was virulent enough to kill mice and the R-strain avirulent. When Griffith heated S-strain cells, killing them, and injected them into mice, the mice lived. When he injected both heat-killed S-strain and live R-strain cells together into mice, the mice frequently died. The process is summarized in Fig. 6.1.

In examining the dead mice, Griffith found living S-strain cells in their bodies! To Griffith, the experiment suggested that there must be something in the S-strain heat-killed cells that could convert the R-strain avirulent cells to the lethal form. What was this "something"? Why did it have such a capability? Griffith did not find this something before he died in 1941. In a biology textbook, the authors wrote, "This ability of some chemical substance in the dead bacteria to convert a related strain to a genetically stable, new form was termed transformation. Later, Avery, Macleod, and McCarty of the Rockefeller Institute determined that the 'transforming principle' was in fact DNA" (Villee et al. 1989, p. 298).

Many historians recognize that Griffith did make a discovery,[1] but is it adequate to call his findings from that experiment an experimental discovery? If so, what did he discover? The novel phenomenon of transformation? Does he deserve recognition for that discovery? Does the recognition of his discovery rely on determination of the transforming substance?

In 1892, Hans Driesch shook a sea urchin embryo at the two-cell stage so vigorously that it became two separate cells. He observed that they developed into complete embryos that were normal in configuration but smaller than natural counterparts. He tried to separate cells from the embryo at the four- or eight-cell

[1] A historian of molecular biology described this event: "[I]n 1928, Griffith in London had published a startling discovery" (Judson 1996, p. 18). Another historian of molecular biology wrote: "In 1928, the British physician Fred Griffith discovered the strange phenomenon of transformation" (Morange 1998, p. 31). The evolutionist Ernst Mayr (1982, p. 818) described this event with similar locution: "In 1928 the British bacteriologist F. Griffith discovered that…" The historian of general biology Lois Magner (2002, p. 428) commented on Griffith's results: "In retrospect, it can be said that Griffith has observed genetic transformation, but he probably did not realize that the phenomenon he had discovered involved the transfer of hereditary material."

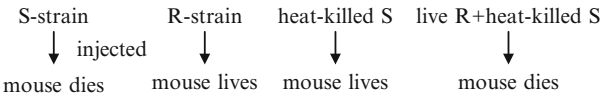


Fig. 6.1 Griffith's experimental procedure and results

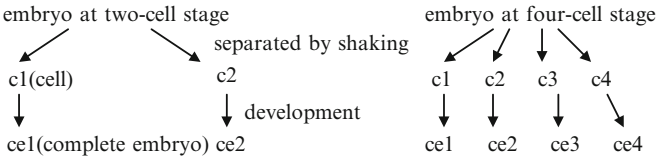


Fig. 6.2 Driesch's experimental procedure and results

stage and observed that all separated cells could develop normally. This result was quite different from Driesch's expectation before the experiment, a surprising phenomenon. Later, he explained it by adopting vitalism and the concept of vital force. Driesch's experimental process is depicted in Fig. 6.2.

This famous experiment conducted by Driesch at the end of the nineteenth century provides an interesting contrast to Griffith's case. Some historians of biology describe Driesch's experiment against the background of Wilhelm Roux's hypothesis of mosaic development and frog-egg experiment,[2] treating Driesch's work as a disproof or anomaly of Roux's "developmental mechanics." Instead of describing Driesch's experiment as a discovery, however, they label his views as "extreme" or involving "confusion."[3]

Here we have a puzzle. Neither Driesch nor Griffith correctly interpreted his experimental results. Why are historians pleased to declare that Griffith discovered a new phenomenon but hesitant to attribute the same achievement to Driesch

[2] Roux and Driesch are regarded as the cofounders of experimental embryology. But they held opposite positions. Roux believed that the development of an embryo is determined by intrinsic factors of eggs in a mechanistic way. He called this "self-differentiation" or "mosaic" development, which means the capacity of the egg or of any part of the embryo to undergo further differentiation independently of extraneous factors or of neighboring parts in the embryo. In other words, parts of an embryo correspond to parts of a developed individual. Thus, Roux called the new discipline employing his view and experimental approach "developmental mechanics." To confirm his beliefs, Roux conducted the famous "pricking experiment," in which he destroyed one of the cells of a frog embryo at the two-cell stage by pricking it with a hot needle. As a result, the undamaged cell developed into a half embryo. See Magner (2002, pp. 195–197).

[3] Mayr (1982, p. 118) wrote: "This unexpected amount of self-regulation induced Driesch, who had performed this experiment, to embrace a rather extreme form of vitalism. . .". Magner (2002, p. 198) commented: ". . .Driesch had apparently reached a more profound level of confusion, which seemed to end all hope of finding a mechanistic explanation for development." Neither did other historians who mentioned Driesch describe his experiment in terms of "discovery" (Carlson 2004; Bowler 1989). Whatever their positions, all have expressed an ambiguous attitude toward Driesch.

(although they respect him as one of the founders of experimental embryology)? Historians have not given us a reasonable account.

In his discussion of "anomaly and the emergence of scientific discoveries," Thomas Kuhn (1970, p. 62) specified three characteristics appearing in the process of scientific discoveries: the previous awareness of anomaly, the emergence of both observational and conceptual recognition, and the consequent change of paradigms. Kuhn claimed that these are characteristics of all discoveries, from which new sorts of phenomena emerge. One may well read the three characteristics as Kuhnian conditions defining scientific discovery. The first condition would be satisfied when an experimental result directed by a paradigm conflicts with the paradigmatic theory and the result is treated as an anomaly by scientists in the normal period; the second, when the anomaly is recognized as significant in both observation and concept; and the third, when a new theory that is able to solve the anomaly is generated.

One may invoke the three conditions to interpret why historians have such a different assessment of the two cases above. Griffith's case satisfies these conditions, because a new theory of hereditary material that could explain the transformation was proposed by Avery and his team[4]; Driesch's case does not satisfy these conditions, because Driesch indulged in the old vitalistic paradigm. The problem is that Griffith himself did not explain his experimental result, although he did not commit to any "old" theory. If a new theory that is able to solve anomalies proposed by other scientists can give the experimenter Griffith credit for his "discovery," then the later theory about *conditional specification* of development should also entitle Driesch to be credited as a discoverer.[5] We should regard Driesch as the discoverer of conditional specification in accordance with Kuhnian definition of discovery, although a "correct" explanation of this phenomenon came much later.

In appealing to Kuhn's view of scientific discovery, however, one should not forget his doctrine of the theory-ladenness of observation. Based on this doctrine, scientists' experimental observations presuppose their precedent paradigmatic beliefs. Thus, Griffith did not observe and interpret the transformation phenomenon, nor Driesch the phenomenon of conditional specification, because they could not correctly interpret their experimental results in accordance with their paradigms. Thus, neither would be entitled to be called a discoverer. Here we see an inconsistency in the Kuhnian view of discovery. According to the conditions of

[4] Judson (1996, p. 18) described the community's response to Griffith's experiment: "It raised clouds of speculative and spurious explanations. . .Avery at first found it impossible to credit Griffith's paper. The findings seemed to overthrow his own fundamental demonstration of the fixity of immunological types. But bacterial transformation was confirmed that same year in Berlin and in 1929 was repeated at the Rockefeller Institute." The description seems to accord with the three characteristics of scientific discoveries specified by Kuhn.

[5] In fact, Scott Gilbert, the author of a textbook of developmental biology, claimed that Driesch "provided the first experimentally observable evidence of conditional specification" (Gilbert 2010, p. 114). The current theory of developmental biology states that "conditional specification is the ability of cells to achieve their respective fates by interactions with other cells" (Gilbert 2010, p. 112).

discovery, both Griffith and Driesch are discoverers. Based on the doctrine of theory-ladenness of observation, however, neither Griffith nor Driesch is a discoverer.

Furthermore, the Kuhnian view suggests that scientific discoveries usually occur over a period of time and involve a number of scientists. Therefore, discoveries should be regarded as collective rather than individual achievements. In this sense, Griffith's experiment was nothing but part of the collective discovery of transformation and the "transforming principle," and Driesch's was part of the collective discovery of conditional specification. Such a view is plausible; however, it seems to reject the concept of experimental discovery and indicates that neither Griffith nor Driesch made a discovery by experiment. Yet it is not clear how much they contributed to the two collective discoveries. From a historical perspective, we want to know how to precisely assess their role in the related discoveries. From a philosophical perspective, we wonder whether there are events or activities that can be qualified as experimental discoveries. The answer to the historical question may rely on the solution to the philosophical one.

3 Historical Controversies About Mendel's Discovery

Almost all textbooks of biology or genetics honor Mendel as the father of classical genetics, because he was regarded as the discoverer of the first two fundamental laws of heredity: *the law of segregation* (or *Mendel's first law*) and *the law of independent assortment* (or *Mendel's second law*). The fist law states that the two copies of a gene segregate (or separate) from each other during transmission from parent to offspring (Brooker 2009, p. 23). The second law states that two different genes will randomly assort their alleles during the formation of haploid cells (Brooker 2009, p. 27). However, almost every textbook provides different formulations for the two laws.[6] Scientists have continued to rewrite them by adding new terms and concepts from later theories of genetics (see Villee et al. 1989; Watson et al. 2004; Hartl and Jones 2005). How was Mendel supposed to discover the two laws by his experiments on plant hybrids?

In the beginning, Mendel thought that "the value and validity of any experiment are determined by the suitability of the means used as well as by the way they are applied" (Mendel 1966, p. 3). The experimental object, pea, was selected because it possesses constant differing traits in the shape of the ripe seeds (round vs. wrinkled), in the position of flowers (axial vs. terminal), in the length of stem (long and short), and so on. Mendel treated each plant having a distinct trait as a form and then

[6] Take the formulations from Villee et al. (1989) as an example. In this textbook, the first law is formulated as "when gametes are formed, the genes behave like particles, becoming separated so that each sex cell (egg or sperm) contains only one member of each pair" (Villee et al. 1989, p. 242). The second law is formulated as "[A]lleles of two or more different loci are distributed randomly with respect to one another during meiosis" (Villee et al. 1989, p. 252).

made the first series of experiments by hybridizing two forms with a pair of contrastive traits. For instance, crossing a long-stem pea plant with a short-stem one would produce the first generation of offspring (the hybrids), in which all plants possess a long stem. Then he made the first generation of offspring self-fertilize and got a ratio of 3 long stem to 1 short stem in the second generation. Mendel called the traits of the majority of the offspring "dominant" and the minority traits "recessive." He used the sign *A* to denote the dominant traits and *a* the recessive; he then expressed the experimental results as a symbolic formula: $A + 2Aa + a$ (Mendel 1966, p. 16). The result has been interpreted as empirical evidence for the first law of heredity.

Mendel conducted the second series of experiments by hybridizing two forms possessing a combination of two pairs of contrastive traits—for example, long stem with axial flowers and short stem with terminal flowers. Using the same procedure as in the first series of experiments, Mendel obtained a second "mathematical formula": $AB + Ab + aB + ab + 2ABb + 2aBb + 2AaB + 2Aab + 4AaBb$, in which *A* denotes the dominant of the first pair of contrastive traits and *B* that of the second pair; *a* denotes the recessive of the first pair and *b* that of the second (Mendel 1966, p. 20). This result has been interpreted as empirical support for the law of independent assortment. Therefore, Mendel was credited with the discovery of the two fundamental laws of heredity.

When he finished his work, Mendel wrote a paper titled "Versuche über Pflanzen-Hybriden (Research on plant hybrids)" to discuss the result and published it in a local journal, *Proceedings of the Natural History Society of Brno,* in 1866. The paper with his "discovery" was neglected for 34 years, until 1900, when three botanists—Hugo de Vries, Carl Correns, and E. Tschermak—rediscovered the laws of heredity and Mendel's paper. Almost all scientists and authors of textbooks believe this story, as do many historians of biology (Mayr 1982; Magner 2002; Carlson 2004). It is the orthodox view of Mendel's discovery. Some historians of biology began to challenge this orthodox view in the 1980s. They attributed Mendel's work to the old tradition of breeding and hybridization experiments rather than the new genetics. They questioned whether Mendel was really a Mendelian, arguing that Mendel had no idea of a gene or paired factors and presented no law of heredity in his 1866 paper. What concerned Mendel was not the problem of how a trait is transmitted from parents to offspring but whether new species could arise by hybridization (Olby 1985; Bowler 1989; Corcos and Monaghan 1993). The historian Bowler (1989) reconstructed the Mendelian revolution based on Kuhn's theory of scientific revolution.[7] Bowler placed Mendel's research and experiments under

[7] Bowler (1989, p. 15) emphasized the importance of the paradigm concept to the history of science: "The fact that paradigms are entities with definite beginnings and ends turns the history of science into a genuinely historical discipline, since it implies that one can only understand the science of a past era by trying to think oneself into the conceptual scheme of the then-dominant paradigm." In Bowler's view, the invention of Mendelian genetics (i.e., the general acceptance of Mendel's laws of heredity) required the construction of a new conceptual framework of "hard heredity." Therefore, Mendel's contemporaries naturally failed to understand the significance of his findings to heredity.

the paradigm of developmentalism, which dominated biological research in the nineteenth century. Developmentalists do not separate the problem of heredity from that of genesis and development. Therefore, developmentalism implies a conceptual framework of "soft heredity," which makes no commitment to the idea of particulate factors.[8] Only after developmentalism was replaced by the shift to the new paradigm of hard heredity at the end of the nineteenth century were scientists, at the beginning of the twentieth century, able to see the hereditary significance of Mendel's experiment in a new light. Mendelian genetics was then built gradually, because the rediscoverers, along with American geneticist Thomas Hunt Morgan's team, reinterpreted Mendel's results based on the new conceptual framework of hard heredity (Bowler 1989: ch. 6). Let me call this alternative to textbook orthodoxy the "paradigm-based account."

If one carefully reads Mendel's paper, one finds that Mendel's original text is more consistent with the paradigm-based account than with the orthodox view. In the first and second paragraphs of the 1866 paper, Mendel clearly stated the background and the goal of his experiment:

> The striking regularity with which the same hybrid forms always reappeared whenever fertilization between like species took place suggested further experiments whose task it was to follow the development of hybrids in their progeny. (Mendel 1966[1866], p. 1)
>
> That no generally applicable law of the formation and development of hybrids has yet been successfully formulated can hardly astonish anyone who is acquainted with the extent of the task... (Mendel 1966[1866], p. 2)

If an experimenter's background and goal guide him to perform an experiment and interpret its result, then certainly Mendel did not perform an experiment on heredity, nor did he interpret the result from the perspective of genetics. In this sense, Mendel did not discover the laws of heredity nor could he be regarded as the founder of classical genetics. After the 1980s, some biologists and historians of biology regarded Mendel as a hybridist, but still insisted that he fully realized the hereditary significance of his experiments; he was therefore entitled to be called the father of classical genetics (Mayr 1982; Hartl and Orel 1992). For instance, Mayr argued that the strongest testimony in Mendel's paper is the word "Elemente" (element), which Mendel postulated to account for the experimental result. Mayr contended,

> He postulated that the characters are represented by "gleichartige [identical] oder differierende [differing] Elemente." He does not specify what these "Elemente" are – who could have done so in 1865? – but considers this concept sufficiently important that he refers to these "Elemente" no less than ten times on pages 41 and 42 of the *Versuche*. Evidently they correspond reasonably well to what we could now call genes. (1982, p. 716)

[8] "Soft heredity" means that transmission of characteristics to the offspring could be modified by changes taking place in the parents' bodies due to new habits or a new environment. In contrast, "hard heredity" rejects the notion of soft heredity and holds that characteristics are transmitted unchanged from one generation to the next. See Bowler (1989, p. 3).

I do not think that "Elemente" correspond well to "genes." Here are several key paragraphs where "Elemente" occurs in Mendel's paper.

> This development proceeds in accord with a constant law based on the material composition and arrangement of the elements that attained a viable union in the cell. (Mendel 1966 [1866], p. 42)
>
> In the formation of these cells all elements present participate in completely free and uniform fashion, and only those that differ separate from each other. In this manner the production of as many kinds of germinal and pollen cells would be possible as there are combinations of potentially formative elements. (Mendel 1966[1866], p. 43)
>
> The distinguishing traits of two plants can, after all, be caused only by differences in the composition and grouping of the elements existing in dynamic interaction in their primordial cells. (Mendel 1966[1866], p. 43)

Although these paragraphs can be interpreted easily in terms of the later Mendelian theory of genetics, no occurrence of the word clearly indicates that those elements are *particulate genes* or *Mendelian factors*. These statements prove only that elements exist, work within cells, are responsible for traits of plants, and are responsible for producing different types of germ cells. It is possible and plausible to interpret Mendel's text as saying that *many* elements are jointly responsible for *a unitary* trait. What is more, in Mendel's discussion about experiments on *Phaseolus* at the end of his 1866 paper, the occurrence of signs A_1 and A_2 is puzzling and might refer to multiple "elements" for one color or color range.[9] So Mendel's use of "element" did not conclusively prove that he had an embryonic idea of the gene.

Darden (1991) reviewed some of the historical literature about Mendel's story, concluding that "a complete consensus has not emerged among Mendel scholars. A means of choosing among competing historical interpretations is often difficult, because the extant historical evidence is insufficient and underdetermines any one account" (Darden 1991, p. 40). Regarding the issue we are concerned with, I think that we could choose a historical account which is most coherent with Mendel's text (see the next section).

In sum, the historical issues concerning Mendel's discovery can be expressed in the following questions:

Q1. Did Mendel discover the two laws of heredity?
Q2. Did Mendel have an embryonic idea of what was later called a Mendelian factor or gene?
Q3. Did Mendel play a key role in the research on heredity?

These questions are interrelated, but they can be answered separately. One will get three affirmative answers according to the orthodox view. Scholars who agree with the paradigm-based account will answer all questions in the negative. Some

[9] In the discussion, Mendel said, "Were blossom color A composed of independent traits $A_1 + A_2 \ldots$, which produce the overall impression of crimson coloration, then, through fertilization with the differing trait of white color a, hybrids associations $A_1a + A_2a + \ldots$would have to be formed" (Mendel 1966, p. 35). This seems to mean that A_1 and A_2 correspond to different elements responsible for one and the same characteristic A.

historians of biology may accept that Mendel did not discover the laws but insist that he had an embryonic idea of the gene and therefore certainly played a key role in classical genetics.

I accept that Mendel did not discover the two laws of heredity, nor did he have an embryonic idea of the gene, but I argue that he definitely played a crucial role in the research on heredity. Still, one may wonder, if Mendel had no embryonic idea of a gene, in what sense can one say that Mendel's experimental work can be regarded as a key? What did Mendel discover if he did not discover the laws? Could his "discovery," whatever it may be, be regarded as an experimental discovery?

4 What Did Mendel Discover?

I think that Alain Corcos and Floyd Monaghan (1993) convincingly analyzed Mendel's discovery in their detailed explication of the entire text of Mendel's "Verschue." They first claimed:

> A close study of his paper reveals that the laws of heredity, which are supposed to be there, are not present. Instead, one finds a series of laws relating to the formation of hybrids, which are entirely different from the traditional "Mendelian" laws of heredity. (Corcos and Monaghan 1993, p. xvi)

The two authors reconstructed the series of five laws from Mendel's original text. However, one may find that they are more like "generalizations" than scientific laws. Given also the fact that the notion of "a law" in biology is often questioned, I will call Corcos and Monaghan's five laws "generalizations." They are:

G1. The hybrid offspring of parents, each true-breeding for one of the contrasting characters of a trait, are all alike and like one of the parents. No intermediate types are formed. (Corcos and Monaghan 1993, p. 81, p. 89, p. 97)

G2. Reciprocal fertilizations yield the same hybrid forms. That is, the hybrid trait will be that of the dominating parent regardless of whether that is the seed parent or the pollen parent. (Corcos and Monaghan, p. 81, p. 89, p. 97)

G3. When the hybrids are allowed to self-fertilize, the offspring always appear in two classes: one class like the hybrids and like one of the original true-breeding parents (the dominating); and one class like the parental character not visible in the hybrid generation (the recessive). No intermediate forms are produced. The two classes occur in the approximate ratio of 3 dominating to 1 recessive. (Corcos and Monaghan, p. 89, p. 97)

G4. (a) When the recessive offspring of the hybrids are allowed to self-fertilize, they always breed true. (b) When the dominating offspring of the hybrids are allowed to self-fertilize, approximately one-third of them breed true while two-thirds of them behave exactly like the hybrid generation. (Corcos and Monaghan, p. 97)

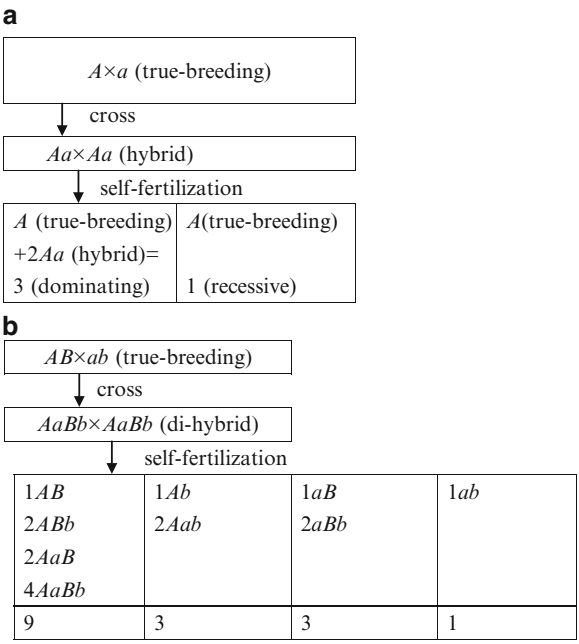


Fig. 6.3 Mendel's first (**a**) and second (**b**) experimental procedures and results

G5. The behavior of each pair of differing traits in a hybrid association is independent of all other differences in the two parents. (G5 is directly extracted from lines 144–146 of Mendel's text. See Corcos and Monaghan, p. 113, p. 118)

One can see that these five generalizations are formulated in terms of traits, true-breeding, hybrids, and other observable objects. No terms from gene theory are used. By contrast, the formulations of the two Mendelian laws of heredity, as later presented in many biology textbooks, include many theoretical terms such as "gene," "allele," "locus," "meiosis," and others from gene theory and cell theory. Because these five generalizations were obtained from Mendel's experiments, one can see that they together describe the two diagrams shown in Fig. 6.3a, b, which represent Mendel's experimental process. In Fig. 6.3a, *A* and *a* refer to the traits representing different forms of true-breeding individuals; *Aa* refers to the trait representing the hybrid from crossing *A* with *a*. In Fig. 6.3b, *A* and *a*, and *B* and *b* refer to a pair of forms of two different traits; *ABb*, *AaB*, and *AaBb* refer to a form of hybrids. *Ab* and *Aab* refer to the second form of hybrids, and *aB* and *aBb* refer to the third one. So the numbers in the four forms of true-breeding and hybrids imply a ratio of 9:3:3:1, but Mendel himself did not express this ratio explicitly.

Figures 6.3a, b can be understood as two data models, because they do not include theoretical or hypothetical terms such as "Mendelian factors," "gene," "locus" (in chromosomes), and the like. Rather, they organize the raw data from Mendel's series of experiments into intelligible phenomena, model the experimental results, and envisage the need to search for an underlying mechanism to be

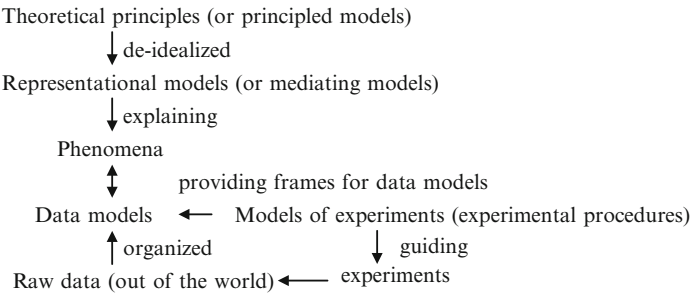


Fig. 6.4 A hierarchical framework for theories, models, experiments, and data

explained. They are also models of experimental data in the sense that they organize and model the data from the experiments. So Mendel really discovered phenomena by proposing two data models—which were expressed in terms of "laws" or "generalizations" in his 1866 paper. Mendel used the generalizations (implying two data models) to describe the patterns of the formation of hybrids. Although Mendel had no notion of a data model, the models described by his generalizations further represent significant phenomena: the regularities of the combination and transmission of traits in some certain crosses between true-breeding forms and hybrids. Therefore, Mendel did make a scientific discovery—this was an *experimental discovery of phenomena*—and Mendel's discovery consisted of *the empirical modeling of experimental data*. Mendel's discovery is a key to classical genetics, because the phenomena and the models are *necessary* for the later "Mendelian" theories of genetics.[10] The view can be confirmed by the fact that the rediscoverers and the ensuing geneticists "invented" the two fundamental laws of heredity and the concept of the gene to explain the genetic phenomena and the models.

What is the notion of a data model, and why is it significant? The notion of a data model has been widely used by philosophers who discuss the role of models in science. They generally agree that there exist models of data that mediate between theories and data, although they have slightly different but largely similar frameworks (Suppes 1962; Mayo 1996; Giere 1999, 2010; Teller 2010; Brading 2010). In order to fit with the goal of this chapter, I adopt the hierarchical framework for theories, models, and data as shown in Fig. 6.4. The framework shows that a model of experimental procedure can guide the performance of an experiment and produce raw data. It can also provide a frame for a data model to organize raw data produced from the experiment. Representational models can be

[10] Mendelian genetics or Mendelian theories of genetics underwent a continuous micro-change from the rediscoverers Hugo de Vries and Carl Correns, to the British Mendelian William Bateson, to the American geneticist Thomas Hunt Morgan. They all have similar but different versions of classical genetics. Darden (1991) impressively and convincingly showed this developmental process of Mendelian genetics.

obtained by de-idealizing principled models and be used to explain phenomena revealed by data models.

By and large, a data model represents the structure of data from observations, measurements, or experiments. Different philosophers have different characterizations of data models.[11] There is no need to develop a general conception of data model here; I intend to give only an account of models of experimental data. A model of experimental data can be produced, constructed, and extracted from a specific experimental design, arrangement, and results (i.e., the whole process), with some conditions for control.[12] Because the experimental process can be represented by a symbolic or conceptual diagram (for instance, the diagrams shown as Figs. 6.1, 6.2, 6.3a, b), the diagram in turn can be seen as a data model. One can also call those diagrams diagrammatical models, each of which represents some specific pattern, structure, or regularity. The diagrammatical model in Fig. 6.1 represents the pattern of effects produced by injecting two different strains (R and S) of bacteria into mice. The diagrammatical model in Fig. 6.3a represents the pattern of effects produced by crossing certain forms of plants possessing pairs of contrastive traits. Therefore, a data model can reveal a significant phenomenon (in Bogen and Woodward's sense; see next section).

Why were the data models of the formation of hybrids in effect a key to the research about heredity? What allowed the models and generalizations to be reinterpreted into the so-called fundamental laws of heredity by later geneticists? From the view of genetics, G1, G2, and G3 can be interpreted as generalizations of trait transmission, a view that presupposes that there are bearers of traits that can move from parents to offspring. In addition, these generalizations also show that dominating traits appear in the offspring of every generation, whereas recessive traits hide in the second generation and reappear in the third generation. One can infer from the phenomenon that the bearers of these traits would not be blended, and thus one can derive the law of segregation (of the trait bearers). G3 and G4 describe the occurrence or nonoccurrence of paired traits and the fixed ratios in the number of offspring. From this one can infer the "law" of dominance—that is, the existence of dominating and recessive bearers—though this is not a general case for most traits. G5 corresponds to the law of independent assortment; it is an empirical expression of the independent distribution of trait bearers. According to such a line of thinking, we should note that "Mendelian" laws of heredity, two or three, are really theoretical rather than empirical, for they imply the notion of *trait bearer*,

[11] Recently, there have been several waves of debate over the relation between data and phenomena, stemming from Jim Bogen and James Woodward's 1988 paper (Bogen and Woodward 1988). For review articles, see Harris (2003), Bogen (2010), Woodward (2010), McAllister (2010), Teller (2010), and Brading (2010).

[12] Mendel set up three control conditions for his experimentation with plants: (1) The experimental plants must necessarily possess constant differing traits, (2) their hybrids must be protected from the influence of all foreign pollen during the flowering period, and (3) there should be no marked disturbances in the fertility of the hybrids and their offspring in successive generations (Mendel 1966[1866], p. 3).

elaborated into the notion of a particulate Mendelian factor or gene. The notion of trait bearer was envisaged by Mendel in his use and speculation of "Elemente," but he did not specify Elemente as particulate, nor did he discover any law of heredity, nor did he develop a theory of genes.

The foregoing discussion shows how Mendel's data models were "grafted" to Mendelian theory of genetics and were interpreted by later geneticists as the empirical evidence for the theory. This process with Mendel's experimental discovery is important, and even crucial, to the discovery of Mendelian mechanism of heredity.

5 Experimental Discovery and Mechanism

Recall Kuhn's views of scientific discovery. Kuhn (1970) argued that the occurrence of anomalous phenomena is usually a prelude to the emergence of new paradigms or theories. It suggests that searching for a plausible explanation of anomalous phenomena should be scientists' main motive for establishing a new theory. In other words, from the perspective of scientific practice, identifying a crucial anomaly seems to play a pivotal role in shaping the pattern of theory changes.

Kuhn's description seems to be consistent with the general process of scientific discoveries. However, for Kuhn, the discovery of a new phenomenon would be recognized only when the anomaly is solved, accompanied by a paradigm shift. In other words, scientists develop a new paradigm to provide a solution for that original anomaly and recognize the solution as a new discovery. On this point I disagree with Kuhn, because the recognition of a new phenomenon usually occurs prior to rather than posterior to the building of new theories. I think that the general process of scientific discoveries would consist of the following stages: the occurrence of novel data, the recognition and discovery of new phenomena, and the building and discovery of new theories, in that order. Mendel's case shows that the experimental discovery of hereditary phenomena led to the construction of Mendelian theories. Now the question is: How can discovery of new phenomena be recognized in the absence of new theories? I have argued, taking Mendel's work as an example, that the establishment of models of experimental data in the recognition of new phenomena is entitled to be called an experimental discovery. However, this is not complete; there is a need to supply a mechanistic condition.

According to the new mechanistic philosophy, a theory can explain a phenomenon by describing its underlying mechanism. Synthesizing my view on the process of scientific discoveries with the mechanistic view of theories, one can see that the experimental discovery of new phenomena is usually a prelude to a scientific discovery—meaning the formation of a new theory and the discovery of mechanisms from the mechanistic perspective. This indicates the key role experimental discovery plays in the discovery of mechanism. If there were no new phenomena to be explained, scientists would have no motive to construct new

hypotheses. Therefore, *producing the need and the motive to discover mechanisms is an important function of experimental discovery*.

Look at Mendel's case. Mendel himself postulated the existence of "elements" to explain his experimental results. The speculation or imagination of elements suggested an underlying mechanism in which *elements* engage in certain activities to produce a fixed ratio in the number of offspring. Yet Mendel did not realize what those elements were. Nonetheless, his speculation, based on his experimental results, envisaged the need to disclose a hidden mechanism, and that need led ensuing biologists to interpret his paper from the view of hard heredity. Explaining data models of the formation of hybrids involves the transmission of traits, because "hybrid" and "true breeding" are defined by the combination and recombination of traits. This indicates that the search for a mechanism is a later step in the process from Mendel's discovery to the development of Mendelian genetics. People can thus recognize Mendel's findings as a genuine experimental discovery that is prior to the discovery of Mendelian mechanism of heredity.

There are various characterizations of mechanisms in the philosophy of science. The one I adopt was proposed by Peter Machamer, Lindley Darden, and Carl Craver (hereafter MDC):

> Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (MDC 2000, p. 3)

In subsequent works, Darden and Craver have articulated the notion of mechanism and have developed a philosophical theory, taking molecular biological and neurobiological cases as examples (Craver and Darden 2001; Craver 2001, 2002, 2005; Darden 2002, 2005, 2006). For the relation between mechanism and phenomenon that is one of the central concerns of this chapter, MDC presented a preliminary connection: "To give a description of a mechanism for a phenomenon is to explain that phenomenon, i.e., to explain how it was produced," and "The organization of these entities and activities determines the ways in which they produce the phenomenon" (MDC 2000, p. 3). As for the meaning of phenomena, Craver and Darden (2001, p. 122) advocated Bogen and Woodward's (1988, p. 317) argument that phenomena should not be confused with data. They agreed that a phenomenon can be understood as follows: "We think of phenomena as relatively stable and repeatable properties or activities that can be produced, manipulated, or detected in a variety of experimental arrangements" (2001, p. 114). Craver and Darden further pointed out that "different experimental arrangements reveal different aspects of the phenomenon" (2001, p. 122). In addition, in discussing "constraints on the organization of mechanisms," Darden and Craver identified "the characterization of phenomena" as a constraint on the search for mechanisms. They concluded (2001, p. 123):

> Characterizing the higher-level phenomenon to be explained is a vital step in the discovery of mechanisms. Characterizing the phenomenon prunes the hypothesis space (since the mechanism must produce the phenomenon) and loosely guides its construction (since certain phenomena are suggestive of possible mechanisms).

I entirely agree with Darden and Craver. Their insight indicates another function of experimental discovery: *constraining the direction for construction of hypotheses to discover mechanisms*.

Experimental discovery has still another function: *organizing data into significant phenomena*. To show this, I want to emphasize that "the phenomenon" to be explained by the description of a mechanism is usually a "significant" phenomenon possessing a repeatable pattern, structure, or regularity. A significant phenomenon is intelligible, raising a why question and a motive for its explanation. It is worth exploring and investigating. But the phenomenon may be hidden and invisible, or not fully revealed, and the data by themselves are not sufficient to reveal it. An experimental discovery makes it present, visible, and intelligible. The discoverer does so by constructing *adequate* and *correct* models to organize fragmentary data, endow data with significance, and reveal hidden structures, patterns, and regularities. Therefore, *organizing data into significant phenomena* or *creating phenomena* in Hacking's sense (Hacking 1983) is really the first and primary function of experimental discovery.

As I have argued, Mendel discovered (or created) the significant phenomenon of trait transmission by implicitly constructing two data models (Fig. 6.3a, b) as described by the five generalizations. Observations on the number of offspring alone would not be enough to constitute a significant phenomenon. This would have amounted only to a large amount of insignificant data about the transmission of traits from parents to offspring, data that would have been fragmentary and unintelligible. Some hybridists before Mendel had obtained Mendelian ratios (Mayr 1982, pp. 648–649), but they were not and should not be regarded as discoverers, for they never proposed any data model or envisaged a mechanism to give significance to their data. (This will be discussed further in the next section.) Thus, I conclude that the relation between experimental discovery and discovery of mechanism can be characterized by the three functions of the former: organizing data, producing a motive to search for a mechanism, and constraining the space of possible mechanisms.

6 What Counts as an Experimental Discovery

Are experimental discoveries recognized without accompanying theories? The answer is an emphatic yes. What conditions are necessary to recognize an experimental result as a discovery? In other words, what conditions define an experimental discovery? In order to synthesize the discussion in the preceding sections, I suggest the following conditions:

(ED1) An experimenter must explicitly or implicitly propose data models to reveal significant phenomena.
(ED2) No established theories can explain the new phenomena.

(ED3) The experimenter must envisage searching for underlying mechanisms for the phenomena, whether or not he or she proposes correct mechanistic explanations.

One may raise a few questions: How are these conditions to be justified? Why is a data model not theoretical? Can experimental discoveries really be recognized without theories?

Because a model of experimental data is the output of an experimental process (a combination of models of experiments and raw data according to the framework in Fig. 6.4), they are not theoretical. They can be constructed without theories. However, one can still question whether designing an experiment does or does not depend on theories—the question of the theory-ladenness of experimentation. It is the case that one can apply theories when designing experiments, but not all experiments are designed based on theories. We should distinguish between the concept of *theory* and that of *background ideas*. All experimental designs require background ideas, but not all background ideas are theoretical. Mendel designed his experiments with peas on the basis of his ideas about the formation and development of hybrids, for which he did not have a complete theory. Furthermore, a data model and a phenomenon allow a variety of theoretical explanations. Mendel himself and Mendelian theorists (including Hugo de Vries, William Bateson, and Thomas Hunt Morgan and his team) proposed different theories to explain the identical ratios, the data models, and the phenomena of trait transmission (see Darden 1991).[13] In this case and other similar ones, phenomena and data models were constructed prior to and independent of any theory.

According to the three conditions, Mendel did make an experimental discovery; it was a key to genetics. Griffith's case also satisfies these conditions, for a data model can be extracted from Griffith's experimental process, and Griffith envisaged searching for an underlying mechanism to explain the *phenomenon of transformation*. Driesch rejected the possibility of searching for a mechanistic explanation, although by his experimental process, he did find a new phenomenon that no established theories could explain. However, his explanation fails to satisfy ED3, so Driesch did not make an experimental discovery. One may question whether this judgment is fair to Driesch, because he did really discover something novel. Answering the question involves a deeper justification for ED3: Why is the envisaging of mechanisms necessary for recognizing an experimental discovery?

The term "discover" means "not to cover," that is, "to remove a cover or shelter over something." When one removes a cover over something, one makes something

[13] David Gooding (1990) provided another example. The phenomenon that the current in a long wire can produce a round magnetic field is extracted from the physicist Jean-Baptiste Biot's experimental process. The contemporary physicists André Ampère, Humphry Davy, Michael Faraday, and Biot himself, respectively, constructed different theories to explain the phenomenon. So a (significant) experimental phenomenon is always revealed by some certain "adequate" and "correct" experimental arrangement. If the experimental arrangement or procedure is inadequate or incorrect, the experimenter may find no significant phenomenon.

visible or transparent. Appropriating Darden's metaphor, to discover is to cause a black or gray box to become a glass (transparent) box. If a scientist's work cannot help herself or others look inside a black or gray box, she has not discovered anything. Driesch turned a black box about the development of embryos into a gray box, but he re-covered it with another facade: the vital force. This prevented people from seeing inside the gray box. So Driesch did not really discover the phenomenon of conditional specification. Admittedly, Driesch did discover something novel, but what he discovered was *irrelevant* to conditional specification, a phenomenon that requires a mechanistic explanation. Driesch identified it as a phenomenon of embryonic development directed by a vital force, so Driesch is the discoverer of that anomalous phenomenon rather than of conditional specification. Of course, Driesch's experiment still made a contribution to the discovery of conditional specification, by providing an anomaly to be solved.

Kuhnians hold that a scientific discovery needs paradigmatic theories to offer conceptual recognition. Thus, neither Mendel nor Griffith nor Driesch is a discoverer. None of them identified or recognized his findings in accordance with a new paradigm, because none proposed such a thing. The Kuhnian holistic view is problematic because it depends on grand paradigms, complete theories, or theoretical principles in the hierarchical framework in Fig. 6.4.

The orthodox view presupposes that a discoverer is the first person who sees a novel phenomenon by observation or experimentation, without conceptual cognition. This explanation produces inconsistent results in certain cases—for instance, Driesch's and Mendel's. Driesch should be the discoverer of conditional specification, for he first produced and observed "that phenomenon." Mendel should not be the discoverer of the Mendelian pattern of trait transmission, because other hybridists had observed it before Mendel did. The orthodox view is problematic because it identifies a discovery only on the basis of pure experience, neglecting the role of conceptual recognition.

My proposal takes a middle way. My judgment of whether Mendel, Driesch, and Griffith, respectively, made scientific discoveries is in agreement with some historians' general view (e.g., Mayr's and Magner's), but I made my judgment for quite different reasons.[14] I agree that a discovery requires a conceptual recognition, but the recognition may not be theoretical; a data model can play the role. A scientist can use the model to recognize that a set of data represents a significant phenomenon—a repeatable pattern or regularity rather than an illusion, trivial appearances, or "mystery experience." However, data models by themselves are not sufficient to complete a discovery. A discoverer should point out a direction that can lead to an advanced discovery. His data model should indicate possibilities for further research, that is, the envisaging of mechanisms. In the sense that ED3 can provide the most coherent explanation for

[14] By contrast, Mayr and Magner did not explain why they regarded Griffith as a discoverer but Driesch as not, given the similar structure of Griffith's and Driesch's experiments.

the cases of Driesch, Griffith, and Mendel—compared with those of other competitors, of Kuhnians, of the advocates of the orthodox view, and of general historians such as Mayr and Magner—ED3 is justified.

7 Conclusion

I have argued that what Mendel found in his experiments is the phenomenon of trait transmission, represented by data models compiled from repetitive modeling of experimental data, and that his experiments are entitled to be called an experimental discovery of a new phenomenon. Instead of being regarded as incomplete, this experimental discovery of the transmission pattern functioned as the basis for the discovery of the mechanisms of Mendelian heredity—which amounts to the construction of classical genetics—by ensuing generations of biologists. One can see that the pattern of experimental discoveries leading to discoveries of mechanisms recurs in many cases of experimental biology—for example, the discovery of chromosomes, of the crossover of chromosomes, of point mutations, of the double helix of DNA, of retrotranscription, and so on.

This chapter concludes by claiming that if an experimental discovery can indeed be justified as a precursor and prerequisite for the later generation of a new theory in experimental biology, it can be adopted by biologists as a strategy for further discovery: Scientists are advised to find data models to represent significant phenomena before constructing theories or discovering mechanisms. This conclusion reveals the independent contribution of experimentation to scientific discovery in addition to the function of testing. It thus can be regarded as a complementary account to Darden's work on building reasoning strategies for the discovery of mechanisms in biology.

Acknowledgments This chapter is revised from the paper presented at Taiwan Conference on the Philosophy of Biology and Economics at National Tsing Hua University in Hsinchu, March 24–25, 2011. I thank the anonymous referees and editors-in-chief for their valuable suggestions and comments on an earlier version of this chapter. I also thank the participants in the conference, Jean-Sebástien Bolduc, Marcel Boumans, Lindley Darden, Alexandre Guay, Till Grüne-Yanoff, Roberta Millstein, James Myers, and David Teira, for their stimulative questions to my conference paper. I especially express my gratitude to Lindley Darden, Roberta Millstein, Szu-Ting Chen, and Hsiang-Ke Chao for their encouragement and suggestions.

References

Bechtel, William. 2006. *Discovering cell mechanisms: The creation of modern cell biology*. Cambridge: Cambridge University Press.

Bechtel, William, and Adele Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton: Princeton University Press.
Bogen, Jim. 2010. Noise in the world. *Philosophy of Science* 77(5): 778–791.
Bogen, Jim, and James Woodward. 1988. Saving the phenomena. *Philosophical Review* 97: 303–352.
Bowler, Peter J. 1989. *The Mendelian revolution: The emergence of hereditarian concepts in modern science and society*. Baltimore: The John Hopkins University Press.
Brading, Katherine. 2010. Autonomous patterns and scientific realism. *Philosophy of Science* 77 (5): 827–839.
Brooker, Robert J. 2009. *Genetics: Analysis & principles*. Boston: McGraw-Hill Higher Education.
Carlson, Elof A. 2004. *Mendel's legacy: The origin of classical genetics*. Cold Springer Harbor: Cold Springer Harbor Laboratory Press.
Corcos, Alain, and Floyd Monaghan. 1993. *Gregor Mendel's experiments on plants hybrids: A guided study*. New Brunswick: Rutgers University Press.
Craver, Carl F. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.
Craver, Carl F. 2002. Interlevel experiments, multilevel mechanisms in the neuroscience of memory. *Philosophy of Science* 69(Suppl.): S83–S97.
Craver, Carl F. 2005. Beyond reduction: Mechanisms, multifield integration, and the unity of neuroscience. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 373–397.
Craver, Carl F., and Lindley Darden. 2001. Discovering mechanisms in neurobiology: The case of spatial memory. In *Theory and method in the neuroscience*, ed. Peter K. Machamer, R. Grush, and Peter McLaughlin, 112–137. Pittsburgh: University of Pittsburgh Press.
Darden, Lindley. 1991. *Theory change in science: Strategies from Mendelian genetics*. Oxford: Oxford University Press.
Darden, Lindley. 2002. Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science* 69(Suppl.): S354–S365.
Darden, Lindley. 2005. Relation among fields: Mendelian, cytological and molecular mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 349–371.
Darden, Lindley. 2006. *Reasoning in biological discoveries: Essays on mechanism, interfield relations, and anomaly resolution*. Cambridge: Cambridge University Press.
Galison, Peter. 1998. *Image and logic*. Chicago: University of Chicago Press.
Giere, Ronald. 1999. Using model to represent reality. In *Model-based reasoning in scientific discovery*, ed. Lorenzo Magnani, Nancy J. Nersessian, and Paul Thagard, 41–58. New York: Kluwer Academic/Plenum Publishers.
Giere, Ronald. 2010. An agent-based conception of models and scientific representation. *Syntheses* 172(2): 269–281.
Gilbert, Scott F. 2010. *Developmental biology*, 9th ed. Sunderland: Sinauer Associates, Inc.
Glennan, Stuart S. 1996. Mechanisms and the nature of causation. *Erkenntnis* 44: 49–71.
Glennan, Stuart S. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69(Suppl.): S342–S353.
Glennan, Stuart S. 2005. Modeling mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 443–464.
Gooding, David. 1990. *Experiment and the making of meaning*. Dordrecht: Kluwer Academic Publishers.
Hacking, Ian. 1983. *Representing and intervening*. Cambridge: University of Cambridge Press.
Harris, Todd. 2003. Data models and the acquisition and manipulation of data. *Philosophy of Science* 70(5): 1508–1517.
Hartl, Daniel L., and Elizabeth W. Jones. 2005. *Genetics: Analysis of genes and genomes*. Boston: Jones and Bartlett publishers, Inc.

Hartl, Daniel L., and Vitezslav Orel. 1992. What did Gregor Mendel think he discovered? *Genetics* 131: 245–253.
Judson, Horace F. 1996. *The eighth day of creation: Makers of the revolution in biology,* 25th anniversary ed. New York: Cold Spring Harbor Laboratory Press.
Kleiner, Scott A. 1993. *The logic of discovery: A theory of the rationality of scientific research*. Dordrecht: Kluwer.
Kuhn, Thomas S. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Machamer, Peter K., L. Darden, and C.F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67(1): 1–25.
Magnani, Lorenzo, N.J. Nersessian, and P. Thagard (eds.). 1999. *Model-based reasoning in scientific discovery*. New York: Kluwer Academic/Plenum Publishers.
Magner, Lois N. 2002. *A history of the life science*. Revised and expanded, 3rd ed. New York: Marcel Dekker, Inc.
Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. Chicago: The University of Chicago Press.
Mayr, Ernest. 1982. *The growth of biological thought*. Cambridge, MA: Harvard University Press.
McAllister, James M. 2010. The ontology of patterns in empirical data. *Philosophy of Science* 77 (5): 804–814.
Mendel, Gregor. 1966. Experiments on plant hybrids. In *The origin of genetics: A Mendel source book*, ed. C. Stern and E. Sherwood, 1–48. San Francisco: W. H. Freeman.
Morange, Michel. 1998. *A history of molecular biology*. Trans. Matthew Cobb. Cambridge, MA: Harvard University Press.
Nickles, Thomas. 1980a. Introductory essay: Scientific discovery and the future of philosophy of science. In *Scientific discovery, logic and rationality*, ed. T. Nickles, 1–59. Dordrecht: Reidel.
Nickles, Thomas (ed.). 1980b. *Scientific discovery, logic and rationality*. Dordrecht: Reidel.
Nickles, Thomas (ed.). 1980c. *Scientific discovery: Case studies*. Dordrecht: Reidel.
Nickles, Thomas. 1987. Methodology, heuristics, and rationality. In *Rational changes in science*, ed. J.C. Pitt and M. Pera, 103–132. Dordrecht: Reidel.
Olby, Robert. 1985. *Origins of Mendelism*. Chicago: University of Chicago Press.
Schaffner, Kenneth F. 1974. Logic of discovery and justification in regulatory genetics. *Studies in History and Philosophy of Science* 4(4): 349–385.
Schaffner, Kenneth F. 1993. *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press.
Suppes, Patrick. 1962. Models of data. In *Logic, methodology and philosophy of science*, ed. Ernest Nagel, Patrick Suppes, and Alfred Tarski, 252–261. Stanford: Stanford University Press.
Teller, Paul. 2010. "Saving the phenomena" today. *Philosophy of Science* 77(5): 815–826.
Villee, Claude A., et al. (eds.). 1989. *Biology*. Philadelphia: Saunders College Publishing.
Waters, C. Kenneth. 2004. What was classical genetics? *Studies in History and Philosophy of Science* 35: 783–809.
Watson, James D., et al. 2004. *Molecular biology of the gene*. New York: Cold Spring Harbor Laboratory Press.
Woodward, James. 2010. Data, phenomena, signal, and noise. *Philosophy of Science* 77(5): 792–803.

Part III
Reconsidering Biological Mechanisms and Causality

Chapter 7
Mechanisms and Laws: Clarifying the Debate

Carl F. Craver and Marie I. Kaiser

Abstract Leuridan (2010) questions whether mechanisms can really replace laws at the heart of our thinking about science. In doing so, he enters a long-standing discussion about the relationship between the mechanistic structures evident in the theories of contemporary biology and the laws of nature privileged especially in traditional empiricist traditions of the philosophy of science (see, e.g., Wimsatt 1974; Bechtel and Abrahamsen 2005; Bogen, Stud Hist Philos Biol Biomed Sci, 36:397–420, 2005; Darden 2006; Glennan, Erkenntnis, 44:49–71, 1996; MDC, Philos Sci, 67:1–25, 2000; Schaffner 1993; Tabery 2004; Weber 2005). In our view, Leuridan misconstrues this discussion. His weak positive claim that mechanistic sciences appeal to generalizations is true but uninteresting. His stronger claim that all causal claims require laws is unsupported by his arguments. Though we proceed by criticizing Leuridan's arguments, our greater purpose is to embellish his arguments in order to show how thinking about mechanisms enriches and transforms old philosophical debates about laws in biology and provides new insights into how generalizations afford prediction, explanation, and control.

C.F. Craver (✉)
Department of Philosophy, Washington University in St. Louis, 1 Brookings Drive, St. Louis, MO 63130, USA
e-mail: ccraver@artsci.wustl.edu

M.I. Kaiser
DFG-Research Group "Causation and Explanation", University of Cologne, Richard-Strauss-Str. 2, D-50931 Cologne, Germany
e-mail: Kaiser.m@uni-koeln.de

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_7,

1 Introduction

Over a decade ago Machamer et al. (2000) suggested that the philosophy of science, especially the biological sciences, could usefully be reconfigured by thinking about how scientists construct, evaluate, and revise their understanding of mechanisms. They boldly asserted that traditional philosophical topics such as causation, discovery, explanation, functions, laws, levels, models, and reduction would be fundamentally transformed by recognizing the centrality to many areas of science of the search for mechanisms. The revolution they envisioned replaced the last vestiges of the once-received positivist gestalt with a new mechanistic vision, expressed in the very language in which scientists talk about their work and sensitive to problems faced within mechanistic research programs in areas as diverse as biology, cognitive science, ecology, and neuroscience.

Though this way of thinking about the philosophy of science has gained rapid and widespread acceptance, it has unsurprisingly attracted a good deal of criticism from those who wonder whether the mechanical philosophy is really as revolutionary as its proponents suggest and from those who think that traditional ways of thinking about the philosophy of science address problems that the mechanical philosophy is ill equipped to handle. And one might be forgiven for thinking that there is no more central battleground in that debate than the perennial issue of the laws of nature. Positivist philosophy of science and its descendents place the concept of a law of nature at the very heart of their thinking about causation, explanation, prediction, and reduction in particular. From that traditional vantage point, it is reasonable to ask precisely how the concept of mechanism, which plays many of the same roles in the new paradigm, is related to the concept of a law of nature.

So conceived, one naturally sees the concept of mechanism as replacement for the concept of laws. And indeed, a casual reading of the mechanistic literature would give the impression that this is precisely what the mechanists intended to do. Mechanists regularly note that the term "law" is descriptively out of place in the biological sciences. Biologists and other scientists of the middle range (neuroscientists, physiologists, psychologists, etc.) seem to avoid the term "law" and conceive of their work instead in terms of the discovery of mechanisms. Furthermore, the mechanist's rejection of a law-centered picture of science is a part of their general rejection of the "Euclidean ideal" (Schaffner 2008) of science, according to which knowledge is arranged in closed deductive axiomatic systems with strict law statements as the axioms. How, they ask, would the philosophy of science look if this formal gestalt, which had already worn quite thin in places, were replaced by a more material, mechanistic, gestalt: one emphasizing the causal structures that scientists much more frequently discuss (see Craver 2002)? Seeing the mechanistic project in this light leads one to ask, as Bert Leuridan (2010) does in a recent paper, whether mechanisms can really replace laws at the heart of our thinking about science.

Leuridan believes they cannot. In this chapter, we assess his arguments. Though his arguments, as we show below, leave one with no compelling reason to maintain the traditional view, his discussion demonstrates the need for greater clarity about the place of laws in mechanistic sciences.

But first the ground rule: All parties to this discussion, as Leuridan points out, agree that the traditional notion of a "strict law," the universally quantified material conditional with unrestricted scope and a good deal besides, has little application in biology and other special sciences. Mechanists have openly embraced a number of arguments for this conclusion, most notably John Beatty's (1995) suggestion that the laws of biology are evolutionarily contingent and Stuart Glennan's (1996, 2002) idea that the generalizations of biology are mechanistically fragile and so probabilistic and prone to breakdown. Other mechanists emphasize that theoretical claims in biology are typically limited in scope, applying only to some species and strains (cf. Hull 1978), and that the scope of such generalizations is restricted to life on earth in a particular epoch (cf. Smart 1963). Whatever the reason, mechanists have been happy to echo these criticisms as evidence of the limited applicability of the traditional law-based view to the philosophy of biology. But contra Leuridan's suggestion, it should also be noted that the mechanist's general opposition to strict laws does *not* entail opposition to the idea that biologists and other scientists of the middle range seek to learn about and describe general facts. None of these arguments showing that the idea of a strict law distorts crucial features of biology shows that there are no general facts about biology or that generalizations play no important role in biological research practice. And no mechanist has ever made such claims.

Though we proceed by criticizing Leuridan's arguments, we have a larger purpose, namely, to illustrate how thinking about mechanisms enriches and transforms the philosophical debate about the role of laws in biology. In our view, the debate over whether or not there are laws in biology has outlived its usefulness. Nobody anymore denies that there are stable regularities that afford prediction, explanation, and control of biological phenomena. Whether such stable regularities count as laws depends on what one requires of laws, but it is undeniable that generalizations of this sort do many kinds of work in biology. What remains is the admittedly difficult work of showing how this is possible. If one takes the biological sciences to be largely dedicated to the search for mechanisms, in contrast, one can begin to ask in relatively precise ways how generalization contributes to the search for mechanisms and, conversely, what the idea of mechanism brings to long-standing questions about how generalizations afford prediction, explanation, and control.

We begin by clarifying Leuridan's thesis and his central ontological and epistemological arguments (Sect. 2). In Sect. 3 we consider Leuridan's ontological claims and argue that Leuridan fails to show that mechanisms must involve regularities (Sects. 3.1 and 3.2) or that there must be fundamental laws without underlying mechanisms (Sect. 3.3). Despite the emphasis Leuridan places on the notion of projection (i.e., extrapolation), he fails to explain why the generalizations of biology are stable and why certain facts can be extrapolated while others

cannot (Sect. 3.1). We show further how the mechanistic perspective provides new resources to ameliorate these extrapolation problems. In Sect. 4 we turn to the epistemological issues. We reject Leuridan's claim that mechanistic models must contain law statements, and we show how mechanistic knowledge contributes to the search for stable generalizations. We conclude that continued debates over whether mechanisms can replace generalizations are likely to be unproductive. We conclude, second, that by taking a mechanistic stance, one gains a new vantage point on old problems about laws and a view to new problems about the construction, evaluation, and revision of models of biological mechanisms.

2 Leuridan's Thesis

In his title, Leuridan asks, "Can mechanisms really replace laws of nature?" He answers, "No." In fact, Leuridan's positive thesis is much weaker than this title suggests.

Before formulating this weaker claim, it is necessary first to clear up some terminology. Leuridan defines laws as "generalization[s] describing a regularity, not some metaphysical entity that produces or is responsible for that regularity" (2010, fn 1). This definition ignores three traditional distinctions that have brought much-needed clarity to the discussions of laws in the philosophy of science. First, we distinguish laws (metaphysical entities that produce or are responsible for regularities) and law statements (descriptions of laws). If one does not respect this distinction, one runs the risk (as Leuridan does) of unintentionally suggesting that sentences, equations, or models are responsible for the fact that certain stable regularities hold. In like fashion, we distinguish regularities, which are statistical patterns of dependence and independence among magnitudes, from generalizations, which describe regularities. Finally, we distinguish regularities from laws, which produce or otherwise explain the patterns of dependence and independence among magnitudes (or so one might hold).

Let us now reconstruct Leuridan's real thesis. First, Leuridan endorses the ground rule of our discussion. Strict law statements, as Leuridan understands them, are nonvacuous, universally quantified, and exceptionless statements that are unlimited in scope, apply in all times and places, and contain only purely qualitative predicates (2010, p. 318). Noting that few law statements in any science live up to these standards, Leuridan argues that the focus on strict law statements (and presumably also on strict laws) is unhelpful for understanding science. Instead, he focuses on the concept of a pragmatic law (or p-law). Following Sandra Mitchell (1997, 2000, 2003, 2009), Leuridan understands p-law statements as descriptions of stable and strong regularities that can be used to predict, explain, and manipulate phenomena. A regularity is stable in proportion to the range of conditions under which it continues to hold and to the size of the space-time region in which it holds (2010, p. 325). A regularity is strong if it is deterministic or frequent. p-law statements need not satisfy the criteria for strict law statements. Thus, Leuridan's

question is not whether mechanisms can replace laws, simpliciter. Rather it is whether mechanisms can replace p-laws and, correlatively, whether descriptions of mechanisms can replace p-law statements in our thinking about science.

Yet Leuridan's thesis is narrower still. He distinguishes two "kinds" of mechanism: complex system mechanisms (cs-mechanisms) and Salmon/Railton mechanisms. Leuridan characterizes cs-mechanisms as stable configurations of robust objects that produce stable behaviors (2010, p. 319; see also Glennan 2002, pp. 344–46).[1] Leuridan does not define Salmon/Railton mechanisms, except to say that they involve causal processes and causal interactions (2010, p. 319). However, if we follow Glennan, they might be understood as "sequences of interconnected events" or "a chain or web of events leading to a particular event" such as "a boy hit a baseball; the baseball ricocheted off the tree and crashed into the window" (Glennan 2002, p. 345). Salmon/Railton mechanisms are singular causal chains. Substituting into Leuridan's title question yields something closer to the question he in fact addresses: "Can cs-mechanisms really replace p-laws?" Leuridan is not always clear to distinguish this ontological question from its epistemological twin: "Can models of cs-mechanisms replace p-law statements?"[2] But in either case, he concludes they cannot.

More precisely, Leuridan presents four theses, two of which he describes as ontological, and two of which he describes as epistemological:

> First, [cs-] mechanisms are ontologically dependent on stable regularities. There are no [cs-]mechanisms without both macrolevel and microlevel stable regularities. [L1]
>
> Second, there may be stable regularities without any underlying [cs-]mechanism. [L2]
>
> Third, models of [cs-]mechanisms are epistemologically dependent on pragmatic laws. To adequately model a [cs-]mechanism, one has to incorporate pragmatic laws. [L3].
>
> Finally, pragmatic laws are themselves not epistemologically dependent on mechanistic models. They need not always refer to a mechanism underlying the regularity at hand. [L4] (Leuridan 2010, pp. 318–19)

We have inserted the qualification to cs-mechanisms specifically, given that Leuridan offers no argument even purporting to show that non-cs-mechanisms are dependent upon p-laws (see Bogen 2005, 2008), a gap to which we return below. Leuridan argues that cs-mechanisms cannot replace p-laws in our thinking about the ontology of science because cs-mechanisms are ontologically dependent on p-laws (L1), but the opposite is not the case (L2). Furthermore, he claims that cs-mechanisms cannot replace p-laws in our thinking about epistemology because models of cs-mechanisms are epistemically dependent on p-laws (L3) and not *vice versa* (L4).

---

[1] The full passage is "Contrary to Salmon/Dowe mechanisms, complex systems mechanisms (cs-mechanisms) are robust and stable. They form stable configurations of robust objects, and as a whole they have stable dispositions: the overall behaviors of these mechanisms" (Leuridan 2010, p. 319).

[2] On closer inspection, Leuridan's question is still too imprecise since it does not specify the purpose for which cs-mechanisms are intended to replace p-laws (or vice versa). Mechanisms and stable generalizations serve many functions in our thinking about science. Perhaps mechanisms are useful for some philosophical purposes and laws are useful for others.

Before we address Leuridan's arguments, it is necessary first to set the record straight. When Leuridan asks "Can cs-mechanisms really replace p-laws?" the word "really" suggests that somebody has claimed that they can. Is this true? Do mechanists really insist that scientists can discover, explain, predict, and control the action potential, heredity, long-term potentiation, natural selection, and neurotransmitter release (to name a just few of the lengthy examples that mechanists have discussed) without forming generalizations about them? Do mechanists think that the Hodgkin-Huxley model of the action potential, the theory of evolution by natural selection, and the current models of long-term potentiation and neurotransmitter release make no appeal to regular occurrences? In a word, no. They are quite explicit on this matter (see, for instance, Bechtel and Richardson 2010, p. 232; Glennan 1996, p. 52, 2002, p. 345; Machamer et al. 2000, p. 3, p. 7; Bechtel and Abrahamsen 2005, fn. 1, 437; Craver 2007, Ch. 3, pp. 233–34). James Bogen (2005), the mechanist most critical to the role of generalizations and regularities in our thinking about causation, stresses at great length the importance of Mitchell's treatment of p-law statements (and the regularities they describe). He also emphasizes the many epistemic roles that generalizations play in the search for mechanisms (cf. Bogen 2005, p. 401):

(a) to describe the phenomenon to be explained;
(b) to suggest and sharpen questions about causal mechanisms;
(c) to describe constraints on acceptable mechanistic models;
(d) to measure or calculate quantities relevant to the mechanism;
(e) to support inductive inferences without which mechanisms could not successfully be studied;
(f) to support extrapolation of mechanistic knowledge to new cases;
(g) to design effective experiments to test mechanisms;
(h) to simulate the behavior of mechanisms.

The list could no doubt go on. In short, no mechanist denies that biologists search for regularities and routinely formulate generalizations (p-law statements) that can be used for prediction, explanation, and control of phenomena. Indeed, it is hard to see how any significant human activity could be pursued without discovering and representing (in some sense) such regularities. The mechanist claims simply that it is useful to ask further about the material structures those generalizations describe and about how this affects the various tasks scientists perform. In many areas of science, scientists seek to describe mechanisms in order to explain, predict, and control phenomena. If one places the idea of mechanism at the center of one's thinking about those sciences, one suddenly sees p-laws in a new light, with new roles to play (compare Bogen's list to Leuridan's emphasis on prediction, explanation, and control). The question is not whether biological phenomena operate in accordance with p-laws or exhibit p-regularities but rather how the search for those regularities fits into the central aim of describing mechanisms.

3 The Ontology of Mechanisms

Let us now consider Leuridan's argument that p-laws are ontologically fundamental to cs-mechanisms. He argues for two component theses (2010, p. 329):

(a) There can be no cs-mechanism without some stable behavior produced by that mechanism (Leuridan calls this the "macrolevel regularity").
(b) There can be no cs-mechanism without some regular behaviors, operations, or activities displayed by or engaged in by the mechanism's component parts (Leuridan calls these "microlevel regularities").

Given Leuridan's definition of a cs-mechanism (cf. 2010, p. 319; see also Sect. 2), (a) is a tautology: "There can be no stable configurations of robust objects that produce stable behaviors without some stable behavior produced by that mechanism." We agree.

Surprisingly, Leuridan offers historical evidence to shore up his case. "In the life sciences," he writes, "reference to mechanisms cannot be detached from matters of projectability" (2010, p. 329). For example, he notes, Thomas H. Morgan intended his work on the mechanisms of heredity in *Drosophila* (see Morgan et al. 1915), work summarized in a book aptly titled *The Mechanisms of Mendelian Heredity*, to apply outside of the laboratory and to other organisms as well.

Mechanists deny neither that Morgan sought projectable generalizations nor that he succeeded in finding them. However, a mechanist might well insist that scientists sometimes seek details about a particular causal mechanism without any interest in generalizing to other cases. Evolutionary biologists might describe the mechanisms that increased the prevalence of a single adaptive trait in a population or that produced a single speciation event. Epidemiologists might be interested in how AIDS first came to South Korea. Archeologists might be interested in the origins of maize cultivation in North America. Ecologists might be interested in the mechanisms causing fish populations to dwindle in the Chesapeake Bay. Morgan was interested in generality, we grant, but sometimes scientists just want to know how some particular event came to pass, and so they describe the particular mechanism that is responsible.

Bogen (2005, 2008) argues persuasively that the concept of causation (understood in terms of "causally productive activities" (2008, p. 112)) and the concept of regularity can (and do) come apart from one another.[3] We have no difficulty imagining quite irregular mechanisms, such as the mechanisms of neurotransmitter release, that work roughly 10 % of the time or a rusty chainsaw that starts arbitrarily infrequently. What matters for the existence of a mechanism is not how frequently it runs in the limit but how it works when it works, however infrequently. Viewed from this perspective, singular, unrepeated causal chains (so-called one-off

[3] Leuridan mentions Bogen's work but rejects it summarily on the ground that Bogen's criticism of regularism relies on a strict (i.e., universal) notion of "regularity" (see 2010, p. 330). But this is false. Bogen frames his discussion in terms of Mitchell's view of generalizations.

mechanisms or Salmon/Railton mechanisms) are a special, limiting case of cs-mechanisms, not something altogether different. While Leuridan's thesis that there can be no *cs-mechanisms* without some stable behavior produced by that mechanism (cf. 2010, p. 330) is tautologically true, Leuridan's unqualified thesis that "there are no *mechanisms* without… macrolevel… stable regularities" (2010, p. 318; our emphasis) is clearly false. One-off mechanisms are mechanisms *without* a macrolevel regularity. So much for the ontological claim.

3.1 Extrapolation of Generalizations

Things look a bit more promising if we reconstruct Leuridan's projectability thesis as a purely *epistemic* thesis. Morgan wanted to apply what he learned about the mechanisms of heredity by studying *Drosophila* in the lab both to flies outside the lab and to other species. Surely the mechanist owes some kind of story about how this is possible. The clear solution, one might think, is to recognize that there are laws – however exception-ridden, probabilistic, and mechanistically fragile – that license this application. And one might insist that Morgan referred to, and indeed formulated, Mendel's second law while making a career of discovering exceptions to independent assortment (see Allen 1978; Darden 1991). Scientists form generalizations, and then they use those generalizations to say what will happen in new cases. Of course, no mechanist denies that induction and extrapolation (or projection) are important to science. But how are p-laws supposed to help with this task? If p-laws are merely law *statements*, as Leuridan defines them, then they are clearly not the kind of thing that can explain why a given regularity is stable and strong. Law statements *express that*, but do not *explain why*, certain regularities are stable and strong. It seems we must understand Leuridan to mean that stable p-regularities themselves (rather than descriptions of p-regularities) are necessary for one to extrapolate mechanistic knowledge. Here, in full, is Leuridan's discussion of stability: "What are the conditions on which the regularity under study is contingent? How spatiotemporally stable are these conditions? And what is the relationship between the regularity and its conditions (is it deterministic, probabilistic, etc.?)" (2010, p. 325). Given that stability is defined in terms of the range of circumstances in which a generalization holds, the epistemological thesis that extrapolation to conditions outside of the laboratory and to conditions in other organisms requires p-regularities, again, amounts to a tautology: if the regularities discovered about *Drosophila* in the laboratory are to hold outside of the laboratory and for other organisms, then there must be organisms outside of the laboratory for which the regularity holds. If a regularity holds only in Morgan's laboratory or only for *Drosophila*, then there is nothing outside of the laboratory or in other organisms about which to extrapolate. But this is not an explanation of why knowledge extrapolates beyond the laboratory; it is simply a claim that it does extrapolate outside the laboratory. Put this point another way: by helping himself to the idea of p-laws, which are by definition stable regularities, Leuridan does nothing to *explain*

why certain facts can be extrapolated and others cannot. Nor does he tell us how to discern which features of a system can be extrapolated from those that cannot. Rather, by invoking the idea of a p-law, he merely asserts *that* there is a distinction between knowledge that can be extrapolated and knowledge that cannot be extrapolated. But the bald statement that there is a difference between the predicates that project and those that do not, conditions that project and those that do not, and times when the consequent really ought to obtain and times when it should not is not a *victory* for p-laws but simply an *assertion* that a central problem for any theory of p-laws has a solution.

Let's push a bit deeper. For a defender of strict laws, which by definition apply always, without exception, and without limitation of scope, it is reasonably clear how knowledge of the laws would warrant extrapolation. For a defender of a robust metaphysical notion of a law, where a law is part of the structure of the world that explains (rather than merely describes) the p-regularities we observe, then knowledge of the laws would presumably warrant extrapolation. But Leuridan weakens the notion of a law so that p-laws are mere regularities and p-law statements are descriptions of these regularities; further, such descriptions are nonuniversal, have exceptions, and apply only in restricted regions of space-time. In effect, he turns p-laws into imperfect regularities with no robust metaphysical backing. Whether such a weakened p-law warrants extrapolation outside of the laboratory depends upon whether one in fact finds that the regularity continues to hold outside of the laboratory, whether the necessary background conditions hold, whether the target instance under consideration is one of the exceptions, or whether it is not. p-laws, as Leuridan understands them, might not warrant extrapolation. The laws might hold only in Morgan's laboratory, after all. At the very least, if one believes that p-laws offer a solution to the problem of extrapolation, then one owes a further story about how one knows when the conditions for extrapolating the regularity have been met. Leuridan offers no such story.

Bechtel and Abrahamsen (2005), whom Leuridan picks out for particular criticism on this matter, argue on independent grounds that it is philosophically unfruitful to think about the problem of generalization (or extrapolation, in Leuridan's vocabulary)[4] in terms of laws. In a section of their paper called "Generalizing without Laws," they criticize law-based views of generalization and develop an alternative, prototype-based account. Because Leuridan does not mention these arguments, we repeat them here. They argue that if one thinks of biologists as attempting to build law statements, paradigmatically represented in terms of material conditionals, then it is difficult to understand the prototypical structure of biological theories. One is tempted to think of biologists as constructing, for example, a law statement of heredity (such as Mendel's laws). When one encounters variation in that mechanism (as Morgan did), one is tempted

[4] One might distinguish generalization (i.e., expanding the scope of the schema within a species/class) from extrapolation (i.e., expanding the scope of the schema beyond the species/class). Given that the parties to this dispute do not draw this distinction, we treat them as synonymous.

to package the variation into the antecedent of the conditional. In fact, however, one finds that biologists typically characterize a mechanism in a particular strain of a particular species (such as wild-type *Drosophila*) and then recognize that there will be subtle variations on that mechanism in other strains, mutants, and species. They are not looking for general law statements that cover all of them but rather for sets of prototypical models that stand in family resemblance relations to one another (cf. Schaffner 1993). To push Bechtel and Abrahamsen's point one step further, prototype models need not be general descriptions. Bechtel and Abrahamsen also call their view an "exemplar" (2005, p. 438) account, noting that models of mechanisms often describe a particular, exemplary case. (Open a biology textbook and look at some diagrams of mechanisms; more often than not, they are cartoons of a single representative mechanism.) On such an exemplar view, generalization is extrinsic to the mechanistic models (exemplars, prototypes); that is, the model need not contain general statements or general representations at all.[5] Leuridan's insistence that the model must contain such things is simply the imposition of a philosophical prejudice onto actual scientific models that have the capacity rather to surprise us if only we open our eyes to them. The generality of such a mechanistic model is a matter of its scope of application and not something that must be represented within the model itself. If one attempts to put the generality in the model itself, to return to Bechtel and Abrahamsen's point, the model has difficulty accommodating the variability characteristic of biological mechanisms.

Curiously, Leuridan fails to consider the possibility that most p-regularities are stable and strong because they are produced or maintained by mechanisms (see, e.g., Bechtel 2009; Craver 2007; Darden and Craver 2002; Glennan 2010; Steel 2008; Wimsatt 1998). Why might Morgan have expected the apparent exceptions to Mendelian heredity he discovered in his lab to apply outside of the lab and in other organisms? The simple answer is this: he expected the mechanisms of heredity outside the lab and in other organisms to be more or less similar to the hereditary mechanisms at work in his *Drosophila*. The p-laws of heredity are stable and strong precisely because there is an underlying mechanism (e.g., involving crossing over and replication of chromosomes) that explains them.

In his book *Across the Boundaries* (2008), Daniel Steel builds on early suggestions by Darden and Craver (2002) to develop an elaborate analysis of how one can extrapolate scientific knowledge based on an understanding of the relevant mechanisms. The idea behind his "comparative process tracing" (Steel 2008, p. 85) approach is simple and helpful: First, one uses a variety of strategies to learn about the mechanism in the model organism. Second, one compares the mechanism in the model organism to the mechanism of the extrapolational target at certain key junctures. That is, one compares the two mechanisms at stages at which the mechanisms are most likely to differ significantly from one another. The fewer significant differences one discovers at these key points, the stronger is the basis for

[5] This view fits with the semantic view of theories that Bechtel and Richardson embrace (cf. 2010, p. 232).

the extrapolation. Crucially, one need not compare all of the entities, activities, and organizational features of a mechanism to those in the target organism in order to assess the likelihood that one's extrapolation will work: one might, for example, compare downstream (rather than upstream) portions of a mechanism, given that crucial differences downstream will indicate crucial differences earlier. Conversely, similarity at a key bottleneck point in the mechanism might allow one to neglect any differences upstream in the mechanism to focus on what comes later (see Steel 2008, p. 90). Furthermore, if one is interested simply in gross qualitative differences, such as whether a given drug is positively relevant for a side effect or negatively relevant for a side effect, certain minute and highly specific differences in the mechanisms might be less germane than the simple matter of whether there is a positive (excitatory) or negative (inhibitory) causal or correlational relationship in the model organism. Thinking about underlying mechanisms, in short, provides new tools for assessing when our knowledge is likely to extrapolate and when extrapolation is more precarious.

Steel's strategies rely primarily on considering the mechanisms that underlie a regularity, but one might also justify extrapolation on the basis of antecedent mechanisms, such as the mechanism of natural selection. That is, one might claim that the hereditary mechanisms in *Drosophila* can be expected to apply outside of the laboratory and in other species because hereditary mechanisms are evolutionarily ancient and therefore widely conserved across the tree of life. As Bechtel (2009) argues, this mechanistic fact about the history of life warrants tentative (heuristic) extrapolation about closely related species: they might use the same mechanism, or a mechanism composed of similar entities and activities, or mechanisms with similar organizational structures. And one might expect evolutionarily ancient mechanisms to be more widely conserved, and so more fitting for extrapolation, than are relatively recent adaptations. This kind of heuristic is especially interesting in the present context given that, according to this heuristic, a *singular* mechanism (the one-off mechanism that produced the tree of life as we now know it) warrants extrapolation of p-laws in extant species.

While we admit that these mechanistic contributions to our understanding of extrapolation solve neither the problem of induction nor Goodman's new riddle of induction (1955), we insist they nonetheless have considerably more content than the bare tautology that p-laws warrant extrapolation because they are stable and strong. Indeed, extrapolation is at least often justified by appeal to knowledge of mechanisms. In sum, it appears that our epistemic reformulation of Leuridan's argument runs into a dilemma. Either his claim is a tautology to the effect that mechanisms must be general if one is to form true generalizations about them, or it is a substantive epistemological thesis that extrapolation is possible only if there are p-laws. If the latter, then we have shown how Leuridan begs the question by presuming, rather than showing, that p-laws solve the extrapolation problem and by asserting, rather than defending, the disputed thesis that p-laws are required for extrapolation. Most importantly, however, we have reviewed some of the progress mechanists have made in thinking about the problem of extrapolation. Focusing on mechanisms provides fruitful and substantive ways of thinking about how

generalizations are extrapolated in scientific practice. It is unclear why Leuridan refuses the mechanist's help in addressing the extrapolation problem.

3.2 Do cs-Mechanisms Require Micro-regularities?

Let us move on, then, to the second route (b) by which Leuridan argues that cs-mechanisms are ontologically dependent on stable regularities (L1). Leuridan claims: "There can be no cs-mechanism without some lower-level (c)P-regularities (i.e., the regular behaviors, operations, or activities displayed or engaged in by the mechanism's parts)" (2010, p. 331). A (c)P-regularity is a causal p-law, a p-law that is "invariant under some range of interventions" (2010, p. 328). Leuridan argues for this thesis using a thought experiment. If the behaviors of all of the parts of the mechanism were to behave completely randomly, by which he means that they do what they do as the result of a "completely random internal process," "this would make it very unlikely to produce a macro-p-regularity, let alone a (c)P regularity" (2010, p. 331).[6] What shall we make of this argument?

Clearly, Leuridan's thought experiment does not support the ontological conclusion that there can be no cs-mechanisms without some p-regularities among the parts. At most, it supports a probabilistic conclusion that cs-mechanisms are unlikely without p-regularities, and such an argument cannot support the negated existential quantifier in Leuridan's second ontological claim (b). The thesis that x is unlikely to have property F is consistent with the claim that x is F and, for nonzero probabilities, entails that x is possibly F (directly contradicting Leuridan's stated thesis). Although randomly behaving components such as those in Leuridan's example would not form a mechanism (given that the behavior of each is causally independent of the behaviors of the others), it is still possible that together they would produce a regularity, even a (c)P-regularity, of some stability and strength. Just how improbable this would be depends upon the number of variables and the number of values they might take. In order to make experimental progress in the discovery of causes and mechanisms, we regularly presume that regularities do not arise merely from chance. However, as the statistics attached to any causal experiment acknowledge, there is always some nonzero probability that the results of the experiment did arise strictly from chance. Now if macro-regularities can obtain even among causally unconnected random events (as in Leuridan's example), then

[6] It should be noted that Leuridan defines "irregularity" in such a way as to effectively exclude discussion of stochastic mechanisms, mechanisms that work only infrequently or whose frequency of operation and stability in space vary over time. A mechanism that works with probability 0.000001 will count as regular on Leuridan's account because one can write a generalization of the form $P(X) = 0.000001$. This is unfortunate as there are a number of interesting questions that one might ask about probabilistic mechanisms and mechanisms whose probability of working varies over time (as one might expect in systems that are regulated). Thanks to Jim Bogen for calling this to our attention.

why would one ever suppose that it would be impossible for them to obtain among causally connected random events (whatever sense we can make of that notion)?[7] It seems there is no interpretation of the idea of irregularly behaving components that sustains even the negative existential thesis entailed by Leuridan's modal claim that no (c)P-regularity can be produced by irregularly behaving components.

Perhaps what Leuridan means to claim is that very few of the mechanisms described in biology textbooks explain higher-level (c)P-regularities without appealing to regularly behaving components. If one wants to discuss the kinds of mechanism that biologists typically study, then one must acknowledge that there are true p-generalizations about the components of mechanisms. True enough. But this claim is entirely independent of the ontological thesis that cs-mechanisms depend on lower-level regularities. And no mechanist denies that there are true p-generalizations about the components of mechanisms.

If the argument does not work for cs-mechanisms, it certainly will not work for mechanisms in general (as his title and introduction suggest). That is, it cannot establish, as Leuridan claims, that there can be no mechanisms without microlevel stable regularities. It seems one-off mechanisms (the "Salmon/Railton mechanisms" discussed above) might well work without microlevel stable regularities. Such mechanisms probably would not be so scientifically interesting, and we might never know about them, but they might well exist.

Leuridan might, at this point, have entered a long debate about the regular character of causality. Perhaps he could endorse the view that the components in a mechanism can properly be said to causally interact with one another only if there exists a p-regularity relating events of one type to events of another type. If all mechanisms have interacting parts, and if there can be no interactions among parts without p-regularities, then there can be no mechanisms without p-regularities. That's certainly an ontological thesis, and it's one with a grand tradition. It's also a view that some mechanists (such as Bogen 2005, 2008; Machamer 2004; Darden 2006) explicitly challenge.

As we mentioned above, Bogen (2005, 2008) argues that causation and regularity are conceptually distinct. One set of Bogen's arguments turns on the implicit thesis that causation is local (or, in other words, intrinsic): that whether A causes B depends on facts about A, B, and their relation to one another and does not depend on how other A-type things and B-type things behave when they interact. What matters instead is whether A and B are connected by some determinate sort of activity. One need not buy the metaphysics of activities to appreciate the intuitive pull of locality. Imagine a world composed only of two billiard balls traveling through space-time toward one another until one day they clack together and fly back in the directions whence they came. Whether they interacted would seem not to depend on whether any other billiard balls ever meet or on whether the same billiard balls ever meet again; neither is true in the world we are considering. The causal interaction is a fact about them and them alone (i.e., an intrinsic feature of

[7] Again, note that Leuridan is operating with a most unorthodox notion of "irregularity."

their interaction); nobody else matters, and so no p-regularity (or any regularity) matters, to whether they interact.

A second kind of argument for the separability of regularity and causation turns on the possibility of causal relations that have no echo in the correlational structure of the world. For example, one might have a mutation that reduces the overall chance that one will get lung cancer (i.e., the mutation has *negative* statistical relevance for cancer) but that, in a few unfortunate individuals, is, in fact, the trigger for lung cancer. And one might get lung cancer in virtue of having that mutation. One might smoke three packs a day (raising the chance of getting lung cancer) and in fact get lung cancer because of the mutation. The actual causal structure in such cases would appear to run counter to the regularities. To borrow a kind of example first described by Jonathan Schaffer (2000), we might imagine two neurons, A and B, synapsing on a third neuron, C. Suppose we know from experimental investigation that the probability of C's firing given A's firing alone is 0.5, that the probability of C's firing given B's firing alone is 0.5, and that the probability of spontaneous firing in C is 0. Now suppose that A, B, and C all fire. These facts leave the causal facts under-determined. For in this situation, it might be that A caused C to fire, that B caused C to fire, or that both A and B caused C to fire. The difference between these possibilities cannot, ex hypothesi, depend on the regularities involved. It would seem that there is a further fact about the actual causal structure of the situation. Regularities, it might be thought, provide evidence about the causal structure of a mechanism. But the causal structure of the mechanism is something over and above the regularities by which that structure can be detected.

We do not insist on the view that causation is intrinsic, actual, and singular. We simply note that Leuridan does not address the heart of the debate about whether regularities are more fundamental than causation and mechanisms. Some philosophers, most explicitly Glennan (2002), Woodward (2002), and Craver (2007), appear to agree (to a first approximation) with the idea that the interactions in a mechanism should be characterized in terms of invariant change relating generalizations. They stress, for example, that knowledge of causes is practically valuable precisely because it is general. And they emphasize the close connection between the generality of causation and the methods used to test causal relations (see Woodward 2004). Bogen, we have seen, disagrees. The merits of and relations among these approaches have been discussed at some length by Craver (2007), Glennan (2002, 2010), Psillos (2004), Tabery (2004), and Woodward (2002, 2010). Leuridan again does not address this discussion.

One last point deserves mention before almost leaving Leuridan's putatively ontological discussion. Leuridan distinguishes between p-laws and (c)p-laws. This distinction is required for Leuridan to distinguish p-laws that are merely useful for prediction from those that, in addition, allow one to explain and control events. One might predict that one is about to run out of gas by looking at one's gas gauge, but the reading on the gas gauge does not explain the emptiness of the tank. Nor could one make it further down the road by breaking the gauge. For this reason, Leuridan (like the mechanists Glennan (2002) and Craver (2007)) appeals to Woodward's

systematic theory of causation (2003). According to that theory, very roughly, causal regularities are stable regularities that continue to hold when one intervenes to change the cause variable. This view of (c)P-regularities, however, depends fundamentally on the idea of an intervention. It also depends on the notion of an ideal intervention, which is one that intervenes *via some causal paths and not others*. It also depends on a thesis of modularity: that it is possible to intervene independently on the different components of a mechanism. As Woodward acknowledges time and again, this view of the semantics of causal claims is not intended as a reductive, metaphysical analysis of the notion of cause. It would be circular as such because one requires an antecedent notion of causation to ground these features of the account (interventions, uncontrolled paths, and modularity). Ironically, a singular notion of causation such as Bogen defends might be just what Woodward's account of intervention and modularity need for their metaphysical ground. If so, then the claim that (c)p-laws are metaphysically more fundamental than singular causation would have the story exactly backwards. But these are complicated matters that we must leave for now.

3.3 Laws Without Mechanisms?

Above we focus on Leuridan's claim that cs-mechanisms ontologically depend on macrolevel (a) as well as on microlevel (b) (c)P-regularities (L1). For the sake of completeness, let us consider Leuridan's second ontological thesis (L2) that there can be (c)P-regularities without underlying mechanisms. Leuridan needs this second thesis to establish the desired "ontological asymmetry between P-regularities and cs-mechanisms" (2010, p. 331). In his hands, this amounts to the claim that it is possible that there are fundamental (c)p-laws, that is, (c)p-laws for which no mechanisms exist. Leuridan does not argue for this thesis, but it seems to us at least conceivable that the world is structured with fundamental (c)p-laws (Glennan 1996, 2002, 2010 embraces this view). To decide whether this conceivable ontological picture is actual, however, would require further argument. It is also conceivable that the world has an infinite series of mechanisms within mechanisms, or that it grounds out ultimately in individual singular causal relations (as Bogen recommends), or perhaps that it grounds out in occurrent matters of fact. Leuridan has no argument to convince us that we are in one of these worlds rather than the other, and we therefore see no compelling reason for a mechanist to take sides.

4 Mechanism and Epistemology

Let us turn finally to Leuridan's claim that p-law statements are epistemically fundamental to mechanistic models. First, he argues that explanatory mechanistic models must include p-law statements (L3), and so mechanistic explanation cannot

proceed in the absence of p-law statements.[8] Second, he claims that mechanistic knowledge is dispensable in our search for p-laws. For instance, by using randomized experimental designs, one can control for disturbing mechanistic factors without knowing what they are. This latter argument is supposed to show that although (explanatory) mechanistic models are "epistemologically dependent" on p-law statements (in the sense that the former require the latter), p-law statements are not "epistemologically dependent" on mechanistic knowledge (L4) (some p-law statements can be discovered without relying on knowledge about mechanisms). Finally, he argues that if our knowledge of laws did depend upon knowledge of mechanisms, then we would face an "infinite (and vicious) epistemological regress" (2010, p. 333). Because knowledge of mechanisms requires knowledge of laws, our knowledge of laws and mechanisms would never ground out in fundamental facts. Fortunately, Leuridan claims, we can know the p-laws without knowing anything about mechanisms, and this blocks the regress.

Leuridan's first point about explanatory models derives from the above discussion (see Sect. 3). If cs-mechanisms require (c)p-laws, then an adequate model of the cs-mechanism requires (c)p-laws. Above, we reject the antecedent. Given that not all mechanisms produce a behavior in a regular way (granted, many do), there exist cases of one-off mechanisms (Salmon/Railton mechanisms) in which the mechanistic model for the irregular behavior necessarily involves neither a macro p-law statement nor a micro p-law statement. Similarly, in cases where a mechanism behaves regularly even though this macro-regularity is sustained by micro-irregularities, a mechanistic model might involve a macro p-law statement but not a micro p-law statement.

But what shall we do about the cases in which biologists explain a general phenomenon in terms of general facts about components and their activities? Mechanists should not, and do not, deny the existence of such explanations. Instead, mechanists deny that an explanatory model *must* be formulated in terms of generalizations.[9] One *would* think that explanatory models of cs-mechanisms must include p-law statements if one embraced a covering law (CL) model of explanation, according to which explanations are arguments that subsume descriptions of events under general law statements. No mechanist, however, accepts the CL model of explanation (see especially Salmon 1984 and Craver 2007). The reasons are too widely known to be repeated here, and it would be

[8] One might have expected Leuridan to defend the epistemic claim that one cannot learn about mechanisms in the absence of p-laws. One might hold that one can test causal connections only on the basis of regularities. Such a claim would be false, of course, as we might make causal inferences on the basis of temporal succession or spatiotemporal contiguity, for example. Leuridan might claim (correctly) that such inferences are fallible, but all inductive inferences are fallible, including those involving p-laws.

[9] Thus, we deviate from the central thesis Holly Andersen (2011) defends in her short response to Leuridan's paper: "The existence of stable regularities in nature is necessary for either model of explanation: regularities are what laws describe and what mechanisms explain" (2011, p. 325).

uncharitable to saddle Leuridan with this much-maligned view when he has not explicitly endorsed a view on the matter.

For mechanists, in contrast, mechanistic explanatory models have explanatory value in virtue of the fact that they represent the relevant portion of the causal structure of the world, not in virtue of the fact that they have a canonical representational form. Explanatory models of mechanisms might be diagrams, equations, exemplars, prototypes, texts, videos, or what have you. In each case, the representational object of the model might be singular or general. What makes these models explanatory, to the extent that they are, is that they correctly (or approximately) describe the causal structures that produce, underlie, or maintain the explanandum phenomenon. Videos and diagrammatic models, for example, need not include *any* linguistic expressions or equations (cf. Perini 2005). Exemplar models, such as a labeled diagram of a single mechanism working from beginning to end, by their very nature describe representative instances rather than general types. The intended scope of the explanatory model, in other words, need not be represented in the model itself. Once one thoroughly abandons the CL model, there is no justification for demanding that explanatory models must include p-law statements. Leuridan's claim that all mechanistic models must include macro p-law statements, that is, generalizations about the behavior of the mechanism, is an un*necessary* restriction on mechanistic models that ignores the plain fact that mechanistic models are frequently developed *without* asserting within the model that it can be generalized to other phenomena. This narrow focus blinds one (as it blinded earlier generations in the philosophy of biology) to the diversity of representational forms one finds in science. Mechanists such as Salmon (1984) and Craver (2007) have therefore rightly separated the question of explanation from the question of how explanatory knowledge is represented.

Let us now consider Leuridan's second epistemic thesis. For the record, we know of no mechanist who insists that one can test p-law statements *only if* one relies on prior mechanistic knowledge. However, let's think through Leuridan's argument. His sole example is a randomized clinical drug trial, in which subjects are randomly sorted into two groups, one of which is given a drug, and one of which is given a placebo. Leuridan perhaps should have acknowledged that the effectiveness of one's randomization procedure might depend upon the mechanism of randomization. What counts as random with respect to one experimental situation will not count as random with respect to another. Were one to survey the extent of homelessness in a geographic region by conducting a poll randomized by street address, the mechanism of randomization would be systematically biased to target people who have homes. Were one to randomize drug trials by zip code, environmental factors could confound the results. The procedure would not be random in the relevant respect. What matters is whether the apparent randomization procedure is likely to sort participants into two groups that have the same distribution of potentially confounding causal factors. Whether the randomization procedure achieves that depends on assumptions about the relevant causal mechanisms at play (even if those assumptions are often so obvious as to be not worth mentioning). Likewise, Leuridan might have acknowledged that the standard procedure of giving

placebos to control groups reflects prior knowledge of the mechanism of the placebo effect. Indeed, experimenters generally take extreme measures to match the experimental and control groups in every way that might possibly confound the results. They test for missing control conditions by asking whether there is some possible difference between the two groups that could plausibly account for the observed changes in the result/effect. Background knowledge about possible mechanisms is often central to that task.

Consider in a bit more detail about how such experiments work. A standard experiment for testing a (c)p-law involves intervening into a putative cause variable, *C*, and detecting from the putative effect variable, *E*. Mechanistic details are often crucial for assessing the appropriateness of one's interventions. As discussed briefly above, one wants to ensure that one's intervention produces the effect in *E* (if any) via *C* and not via some other mechanism. That is, the intervention should change *C*. It should not change *E* directly. It should not change directly the value of any variable between *C* and *E*. Furthermore, the intervention on *C* itself should not be correlated with any other variable that is a cause of *E* (unless it is causally intermediate between *C* and *E*). In some cases, one wants to ensure that the intervention severs the causal influences of other variables on *C* so that one can attribute any change in *E* to the intervention alone. All of these assumptions behind the use of interventions to test (c)p-laws are assumptions about the causal structure, the mechanisms, involved in the intervention technique and in the system under study. An adequate philosophy of experimental intervention thus might make considerable progress by asking how mechanistic knowledge enters into these test procedures (see Woodward 2003; see summary diagram in Craver 2007, Ch. 3).

What about the detection component of a test for a (c)p-law? Allan Franklin (2009) has generated a useful list of strategies by which scientists confirm that their techniques are reliable indicators of phenomena such as *E*. Many of these strategies rely crucially on facts about the mechanisms at play. One might, for example, argue that there could be no other cause of the measured value of *E* besides the fact that *E* has that value. One might show that one's technique reliably registers reliable artifacts known to be produced under aberrant causal conditions. One might rely on a theoretical understanding of the mechanism by which the detection technique works. One might check the results of one's technique against another technique that relies on causally independent mechanisms (see Franklin 2009). In each of these cases, one relies on knowledge about the mechanisms involved in the system and in the detection technique to argue that the methods in question provide an adequate measure of *E* in these circumstances. In short, even if it is possible to test (c)p-laws without knowing the mechanisms (and we deny that Leuridan's example shows as much), one might learn a great deal about how (c)p-laws are tested by thinking about the mechanisms involved in the test conditions. By casting the debate as a forced choice between laws and mechanisms, one occludes far more interesting questions about how mechanistic knowledge contributes to the design and interpretation of experiments for testing p-laws.

Finally, Leuridan claims that if our ability to test (c)p-laws relies exclusively on cs-mechanisms, then we face an infinite regress. The regress arises because if

cs-mechanistic knowledge relies upon knowledge of (c)p-laws (or, more precisely, mechanistic explanations must involve p-law statements), and if knowledge of (c)p-laws requires knowledge of cs-mechanisms, then we never reach the epistemic bottom. It is not clear to us that this is a well-formed problem, and so we are not clear how to solve it.[10] We know of no foundationalist who proposes to build scientific knowledge out of the basic building blocks of mechanisms or laws. Foundationalists tend to construe the epistemic foundation of science in terms of particular matters of fact, sense data, or innate ideas, not in terms of p-laws or mechanisms. We think that both p-laws and mechanisms contribute to the advancement of science, and we feel no pressing need (and have been given no compelling argument) to place one above or below the other in the order of our knowledge. Furthermore, it is not at all clear from Leuridan's formulation how laws stop the regress. If one must know p-laws in order to adequately test p-laws (e.g., p-law statements that one's randomization procedure regularly randomizes, that one's interventions work the same way each time, and so on), then one still has a regress of sorts, and Leuridan has not shown how it will come to an end. How can we design a randomized experiment if we cannot trust that our randomizing procedure generally randomizes? And how can we control for confounding factors if there are no general facts about which factors are confounding? How do we know that our intervention is adequate if there are no general facts about how our intervention works? It would appear that laws are no more epistemically secure than are mechanisms in the foundationalist view that Leuridan apparently embraces.

5 Conclusion

For the discussion of these matters to move forward, it is crucial not to manufacture an artificial conflict between philosophers who emphasize the centrality of mechanisms in our thinking about science and philosophers (such as Mitchell) who seek a plausible way to talk about generalization in science. No mechanist denies that there are pragmatically useful regularities. And nobody who thinks there are pragmatically useful regularities should feel any pressure to deny that the search for mechanisms is central to the practice of biology and many other sciences.

It is a surprising fact about the history of the philosophy of science that of these two correlative concepts, generalizations have tended to dominate the discussion. Against this backdrop, mechanists should be read as suggesting something of a gestalt shift in which mechanisms are moved into the foreground. Such a shift leads attention away from the formal structure of scientific theories (and questions about the logical structure of law statements and models) and toward the material structures that scientists endeavor to describe. Attention to such material structures

[10] Contrary to Leuridan's claim, Machamer et al. (2000) discuss bottom-out activities not as a way of solving some sort of epistemic regress but as a disciplinarily relative way of identifying when explanations come to an end.

provides resources for thinking about how generalizations and mechanisms are discovered, evaluated, and extrapolated and into how such concepts are deployed in explanation, prediction, and control. The perceived need to defend laws, no matter how much they have been weakened and stripped of their once-robust metaphysical content, reflects a conservative refusal to acknowledge that perhaps the philosophy of science might benefit from coming at its subject matter from a fresh perspective. Mechanists decenter laws in their thinking about science because the old paradigm, centering laws, has become mired in debates that are inconsequential and, as a result, have stopped generating new questions and producing new results. In this chapter, we have argued that by trying on the mechanistic gestalt, one can make progress on problems concerning explanation, laws, prediction, and manipulation where the nomic approach seems to have run out of gas. Moving forward, there are far more interesting and better-motivated questions to ask than whether mechanisms can replace generalizations or vice versa.

Acknowledgments We would like to thank Jim Bogen, Lindley Darden, Alexander Reutlinger, and the members of the DFG research group "Causation and Explanation" for comments on earlier drafts. This project was made possible by the funding provided by the Deutsche Forschungsgemeinschaft (DFG).

References

Allen, Garland E. 1978. *Thomas Hunt Morgan: The man and his science*. Princeton: Princeton University Press.

Andersen, Holly K. 2011. Mechanisms, laws, and regularities. *Philosophy of Science* 78: 325–331.

Beatty, John. 1995. The evolutionary contingency thesis. In *Concepts, theories, and rationality in the biological sciences*, ed. Gereon Wolters and James G. Lennox, 45–81. Pittsburgh: University of Pittsburgh Press.

Bechtel, William. 2009. Generalization and discovery by assuming conserved mechanisms: Cross species research on circadian oscillators. *Philosophy of Science* 76: 762–773.

Bechtel, William, and Adele Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Bechtel, William, and Robert C. Richardson. 2010. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge: MIT Press.

Bogen, James. 2005. Regularities and causality; generalizations and causal explanations. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 397–420.

Bogen, James. 2008. Causally productive activities. *Studies in History and Philosophy of Science* 39: 112–123.

Craver, Carl F. 2002. Structures of scientific theories. In *Blackwell guide to the philosophy of science*, ed. P.K. Machamer and M. Silberstein, 55–79. Oxford: Blackwell.

Craver, Carl F. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon.

Darden, Lindley. 1991. *Theory change in science: Strategies from Mendelian genetics*. New York: Oxford University Press.

Darden, Lindley. 2006. *Reasoning in biological discoveries*. Cambridge: Cambridge University Press.

Darden, Lindley, and Carl F. Craver. 2002. Strategies in the interfield discovery of the mechanism of protein synthesis. *Studies in History and Philosophy of Biological and Biomedical Sciences* 33: 1–28.

Franklin, Allan. 2009. Experiment in physics. In *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), ed. Edward N. Zalta. URL: http://plato.stanford.edu/archives/spr2010/entries/physics-experiment/
Glennan, Stuart S. 1996. Mechanism and the nature of causation. *Erkenntnis* 44: 49–71.
Glennan, Stuart S. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69: 342–353.
Glennan, Stuart S. 2010. Mechanisms, causes, and the layered model of the world. *Philosophy and Phenomenological Research* 81: 362–381.
Goodman, Nelson. 1955. *Fact, fiction, and forecast*. Cambridge: Harvard University Press.
Hull, David L. 1978. A matter of individuality. *Philosophy of Science* 45: 335–360.
Leuridan, Bert. 2010. Can mechanisms really replace laws of nature. *Philosophy of Science* 77: 317–340.
Machamer, Peter. 2004. Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science* 18: 27–39.
Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67: 1–25.
Mitchell, Sandra D. 1997. Pragmatic laws. *Philosophy of Science* 64(Proceedings): 468–479.
Mitchell, Sandra D. 2000. Dimensions of scientific law. *Philosophy of Science* 67: 242–265.
Mitchell, Sandra D. 2003. *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.
Mitchell, Sandra D. 2009. *Unsimple truths. Science, complexity, and policy*. Chicago/London: University of Chicago Press.
Morgan, Thomas H., Alfred Sturtevant, Hermann Muller, and Calvin Bridges. 1915. *The mechanism of Mendelian heredity*. New York: Henry Holt & Company.
Perini, Laura. 2005. Explanation in two dimensions: Diagrams and biological explanation. *Biology and Philosophy* 20: 257–269.
Psillos, Stathis. 2004. A glimpse of the *secret connection*: Harmonizing mechanisms with counterfactuals. *Perspectives on Science* 12: 288–319.
Salmon, Wesley. 1984. *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.
Schaffer, Jonathan. 2000. Causation by disconnection. *Philosophy of Science* 67: 285–300.
Schaffner, Kenneth F. 1993. *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press.
Schaffner, Kenneth F. 2008. Theories, models, and equations in biology: The heuristic search for emergent simplifications in neurobiology. *Philosophy of Science* 75: 1008–1021.
Smart, J.J.C. 1963. *Philosophy and scientific realism*. London: Routledge.
Steel, Daniel P. 2008. *Across the boundaries. Extrapolation in biology and social science*. Oxford: Oxford University Press.
Tabery, J. 2004. Synthesizing activities and interactions in the concept of a mechanism. *Philosophy of Science* 71: 1–15.
Weber, M. 2005. *Philosophy of experimental biology*. Cambridge: Cambridge University Press.
Wimsatt, William C. 1998. Simple systems and phylogenetic diversity. *Philosophy of Science* 65: 267–275.
Wimsatt, William C. 1974. Complexity and organization. In *PSA 1972*, Boston studies in the philosophy of science, vol. 2, ed. K.F. Schaffner and R.S. Cohen, 67–86. Dordrecht: Reidel.
Woodward, James. 2002. What is a mechanism? A counterfactual account. *Philosophy of Science* 69: 366–377.
Woodward, James. 2003. *Making things happen. A theory of causal explanation*. Oxford: Oxford University Press.
Woodward, James. 2004. Counterfactuals and causal explanation. *International Studies in the Philosophy of Science* 18: 41–72.
Woodward, James. 2010. Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy* 25: 287–318.

Chapter 8
Natural Selection and Causal Productivity

Roberta L. Millstein

Abstract In the recent philosophical literature, two questions have arisen concerning the status of natural selection: (1) Is it a population-level phenomenon, or is it an organism-level phenomenon? (2) Is it a causal process, or is it a purely statistical summary of lower-level processes? In an earlier work (Millstein, Br J Philos Sci, 57(4):627–653, 2006), I argue that natural selection should be understood as a population-level causal process, rather than a purely statistical population-level summation of lower-level processes or as an organism-level causal process. In a 2009 essay entitled "Productivity, relevance, and natural selection," Stuart Glennan argues in reply that natural selection is produced by causal processes operating at the level of individual organisms, but he maintains that there is no causal productivity at the population level. However, there are, he claims, many population-level properties that are causally relevant to the dynamics of evolutionary processes. Glennan's claims rely on a causal pluralism that holds that there are two types of causes: causal production and causal relevance. Without calling into question Glennan's causal pluralism or his claims concerning the causal relevance of natural selection, I argue that natural selection does in fact exhibit causal production at the population level. It is true that natural selection does not fit with accounts of mechanisms that involve decomposition of wholes into parts, such as Glennan's own. However, it does fit with causal production accounts that do not require decomposition, such as Salmon's Mark Transmission account, given the extent to which populations act as interacting "objects" in the process of natural selection.

R.L. Millstein (✉)
Department of Philosophy, University of California, Davis,
One Shields Avenue, Davis, CA 95616, USA
e-mail: rlmillstein@ucdavis.edu

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_8,

1 Introduction

In the recent philosophical literature, two questions have arisen concerning the status of natural selection: is natural selection a causal process or is it a purely statistical aggregation? And second, is natural selection at the population level or at the level of individual organisms? In an earlier work, I argue that natural selection should be understood as a population-level causal process, rather than a purely statistical population-level summation of lower-level processes or as an organism-level causal process (Millstein 2006).[1]

In reply, Stuart Glennan (2009) argues that (1) natural selection is produced by causal processes operating at the level of individual organisms but that there is no *causal productivity* at the population level and (2) there are many population-level properties that are *causally relevant* to the dynamics of evolutionary processes. In making these replies, Glennan relies on a claim that there are "two types of causes,"[2] causal productivity and causal relevance.

I agree with Glennan's second claim concerning the causal relevance of natural selection at the population level, but I disagree with his first claim concerning the lack of causal productivity of population-level selection processes. Thus, my focus in this chapter will be on the first claim; I will argue that natural selection is *produced* by causal processes operating at the population level.

In what follows, I will first review Glennan's distinction between causal production and causal relevance, followed by an exegesis of his arguments for the claim that there is no causal production at the population level of natural selection. I then respond to each of his arguments. Finally, I offer positive reasons for thinking that there is casual production at the population level of natural selection processes.

[1] In this earlier work (Millstein 2006), I referred to an "individual-level" causal process instead of an "organism-level" causal process. This was a somewhat unfortunate choice of terminology on my part, since, as I will discuss below, populations are themselves individuals. On the other hand, the advantage of that terminology was that it was agnostic with respect to the units of selection; the individuals in question could be genes, cells, organisms, etc. So, to be clear – in this chapter, for the sake of simplicity – I discuss only populations of *organisms*, with the understanding that selection can occur in populations of other entities. The more general question, then, which I will not be discussing here, is whether natural selection consists of causes that act on the individuals *of any sort* that constitute a population (including a population of populations) or whether natural selection consists of causes that act on the population as a whole. Also, in this chapter I will be discussing Salmon's sense of the term "causal process"; what I call a "causal process" in my 2006 paper would probably be, in Salmon's terms, part of a "causal nexus." I will return to this point briefly at the end of this chapter.

[2] Others have also argued for *causal pluralism*, for example, Cartwright (2004) and Hall (2004).

2 Glennan's Causal Pluralism: Causal Productivity and Causal Relevance

Glennan gives the following examples of *causal productivity*:

- The bowling ball knocked over the pin.
- The explosion made Edward deaf.
- The firing of neuron A caused the firing of neuron B.

Causal productivity, according to Glennan, is:

- A relation between events (where an event is an object *doing* something).
- Local – spatiotemporally contiguous or connected by contiguous intermediates.
- Transitive – if A produces B, and B produces C, then A produces C.
- Tied to mechanistic accounts of causation.

It is the connection to mechanistic accounts of causation that most concerns us here. Glennan mentions the following mechanistic accounts as exhibiting causal production: his own (1996, 2002); Salmon's (1984); Machamer, Darden, and Craver's (2000) [hereafter MDC]; and Dowe's (2000). In Glennan's terms, causally connected events require intervening mechanisms involving interacting objects (or parts or components[3]). In MDC's terms, mechanisms consist of entities engaging in activities that produce change. In Salmon's terms, causal processes "are continuous paths of objects through space-time that can interact when they intersect, producing changes in the properties of the objects that constitute those processes" (Glennan 2009, p. 328). Although of course Glennan has defended his own account of mechanisms, for the purposes of his arguments concerning productivity and natural selection, he deems the differences in terminology and detail among the accounts of mechanisms to be not significant.

According to Glennan, *causal relevance* is a counterfactual relation of dependence between a fact *f* and an event *e*. Glennan gives the following examples of causal relevance:

- The fact that Mom did not turn off the hose was causally relevant to her basement flooding.
- The fact that the key has a certain shape is causally relevant to whether it will open the door.
- The fact that the wind is over 30 mph increases the likelihood that a serious fire will occur.

Glennan argues that there are some cases of apparent causation that fit causal relevance but not causal production. In the "Mom" example above (a so-called omission cause), it is true that if Glennan's mother had turned off the hose, her basement would not have flooded; since the counterfactual is satisfied, failure to

[3] See Bechtel and Abrahamsen (2005).

turn off the house is causally relevant to the basement flooding. However, failure to turn off the hose is not an *event* (there is no object *doing* something), and as a result, *locality* is not satisfied, either; thus, the "Mom" example fails to exhibit causal production, according to Glennan. That is, we cannot, Glennan asserts, say that the fact that his mother did not turn off the hose *produced* her flooded basement. On the other hand, Glennan maintains that there are some cases of apparent causation that fit causal production but not causal relevance, such as cases of overdetermination (Glennan 2010b). In overdetermination cases, each putative cause is sufficient to produce the effect, but neither is necessary, so that one cannot say that if the cause had not occurred, the effect would not have occurred (i.e., the counterfactual is not satisfied).

Glennan claims that full understanding of the causal basis of an event requires both the causally productive causes and the causally relevant causes and can be expressed in the form: event *c* causes event *e* in virtue of fact *f*. I myself am not fully convinced that there are two types of causes; indeed, I suspect that accounts of causal relevance and causal production reveal different aspects of the same phenomenon and that there are ways of handling the omission and overdetermination cases. However, as nothing I intend on arguing for in this chapter turns on causal monism, I will assume, for the sake of argument, that causal pluralism of the type that Glennan endorses is true. Moreover, I will mainly focus on causal production, since the question I am examining is whether natural selection exhibits causal production at the population level.

3 Glennan's Arguments Against Population-Level Causal Production in Natural Selection

To try to show that natural selection fails to exhibit causal production at the population level, Glennan gives an example of frequency-dependent selection, which seems like it would exhibit population-level causation if any kind of selection does (Millstein 2006). He asks us to imagine a population of light and dark water bugs whose survival depends on not being seen by a predator fish. The rarer form is always fitter than the more common form because the predator fish form a stereotypic searching image associated with the more common color. Thus, when the light-colored bugs are rarer, they are fitter, but once the light bugs come to predominate in the population, the dark bugs become rarer and thus fitter.

Glennan says that the water bug example shows how and why *the frequency of a color form* (a population-level property) is *causally relevant* to that form's fitness as well as to changes in the distribution of forms within the population (a population-level effect). Indeed, I have argued that natural selection in general (i.e., not just frequency-dependent selection) satisfies counterfactual accounts of causation; if there were no heritable differences in physical characteristics among the organisms in a population (a population-level property), then there would be no differences in reproductive success. In other words, there would be nothing to be selectively

favored or disfavored, as all the organisms would be of the same genotype (Millstein 2006). However, Glennan claims, we cannot strictly say that increased frequency of a form within a population *produces* decreased fitness of that form, because production is a relation involving objects and events, while the population is not (in this case at least) an individual object and the increase of frequency or decrease of fitness are not individual events:

> The only entities here are the fish and the bugs, the only activities are the activities of individual fish and bugs, and the only interactions are when the fish eat the bugs and when the bugs make baby bugs. (Glennan 2009, p. 331)

It is only at the level of the activities and interactions of individual bugs, he argues, that we find the *mechanisms* that produce new bugs.

Glennan's crucial claim here is that populations are not objects (or in MDC's terms, entities) in this example. Glennan gives three reasons for thinking that populations are not objects: (1) entities need to be localized in space and time; they need to engage in particular activities at particular times and places. But, he asserts, the population in the water bug case does not have these properties; the population as a whole is spread out and does not engage in collective activities. The only activities are those of the individual organisms – swimming, evading predators, eating, etc. – and these are not activities of the population as a whole. (2) What makes a collection of parts into a single entity is that these parts have a stable structure, that the stable structure engages in activities as a unified entity, and that these collected parts share a common fate. But, Glennan claims, when a fish kills a water bug, it kills the whole water bug – it cannot kill its legs but not its body. On the other hand, when a fish kills a water bug, it does not kill the whole population of water bugs. The life of one water bug is more or less independent of another. (3) One cannot say categorically that populations either are or are not individual entities; the question of whether they are individuals only makes sense in the context of analyzing a particular causal process. He allows that an ant colony or a baboon troop may be an individual, but in this case the bugs in the pond are not.

Furthermore, according to Glennan, population-level properties do not produce change because *the population is not a part of the mechanism* that produces changes in genotype and phenotype frequencies. On Glennan's account of mechanisms, the parts of the mechanism have to interact with other parts in order to produce the behavior of the whole. But, he asserts, the population as a whole does not interact with other entities as a whole in order to change its genotype and phenotype frequencies.

4 Responses to Glennan's Arguments

It is this last presupposition of Glennan's – that causal production is mechanistic production involving parts and wholes – that I will question first. I will then argue that populations *do* exhibit the characteristics that Glennan says are necessary to be

causally productive. Thus, populations as a whole, at a given point in time, *can* causally produce future states of the same population.

4.1 Non-decompositional Causal Production

Elsewhere, Rob Skipper and I (Skipper and Millstein 2005) argue that natural selection is not a mechanism in Glennan's (or MDC's) senses. So, to some extent Glennan and I agree. However, the problem is not, as Glennan states, that "the population as a whole does not interact with other entities as a whole in order to change its genotype and phenotype frequencies" (Glennan 2009, p. 335). Indeed, there is at least prima facie reason to think that populations of water bugs as a whole often *do* interact with other entities as a whole. For example, a 1969 study of *Sigara distincta* (the organism on which Glennan's water bug example was based) suggests that an increase in water bugs in a particular location was due to an invasion (discussed in Macan 1976). Here "invasion" is not in the sense of an "invasive species," where a few organisms colonize a new area and reproduce rapidly; rather, it is an invasion analogous to that of an invading army. That is, the water bugs migrated as a whole, which undoubtedly changed the genotype and phenotype frequencies in the populations that they migrated from and to. (I give other examples of populations acting as a whole below.) So again, the problem is not that the population as a whole does not interact with other entities as a whole in order to change its genotype and phenotype frequencies.

Rather, one of the reasons that Skipper and I were unable to construe natural selection as a mechanism in Glennan's sense is that, on his account, the interactions among the parts of a mechanism are supposed to explain the behavior of the whole. In other words, mechanistic explanations involve decomposing the whole into its parts (or entities and activities, on the MDC view). However, if it were the case that a population could interact with other entities as a whole to produce changes in the very same population, this would not seem to fit the Glennan and MDC models of mechanistic explanation: the interactions of the whole would be what explain the behavior of the whole. In other words, the explanation would not be decompositional in the way that mechanistic explanations on the Glennan and MDC accounts – instances of what Skipper and I call the "new mechanistic philosophy" – seem to be.

Here it might be objected that the accounts propounded in the new mechanistic philosophy are not, in fact, decompositional.[4] After all, Darden argues that "finding the mechanism for the segregation of genes did not require decomposing genes into their parts, but required finding the wholes, the chromosomes, on which the parts, the genes, ride" – in other words, finding the mechanism required going "up" in size level rather than "down" (Darden 2005; see also Darden 1991). Glennan, for his part, has recently given an example of an *ephemeral* mechanism which occurs "at"

[4] Thanks to Carl Craver, Lindley Darden, and Stuart Glennan for each pushing me on this point.

a level: the death of the French literary critic, Roland Barthes, who was struck by a laundry truck while crossing a Paris street on the way home from meeting with then-President François Mitterrand (Glennan 2010a). Other defenses of the new mechanistic philosophy, such as Bechtel and Abrahamsen (2005), Craver (2007), and Craver and Bechtel (2007), emphasize both the multilevel nature of mechanistic explanation and the importance of situating of a mechanism in its context (see especially Craver 2001 on this latter point). So, how can I claim that accounts under the new mechanistic philosophy are decompositional?

A distinction made by Salmon between *etiological explanations* and *constitutive explanations* is useful in answering this question. Salmon states that both types of explanation are "thoroughly causal." However, according to Salmon, etiological explanations "explain a given fact by showing how it came to be as a result of antecedent events, processes, and conditions" (1984, p. 269). Constitutive explanations, on the other hand, show "that the fact-to-be-explained is constituted by underlying causal mechanisms"; they exhibit "the internal causal structure of the explanandum" (1984, p. 270). I would suggest that etiological explanations are "at" a level, whereas constitutive explanations cite lower levels by citing the parts that make up the whole (i.e., they are decompositional). According to Salmon, we can expect that most explanations will have both etiological aspects and constitutive aspects, but we should also recognize that there are some cases of pure etiological explanation and some cases of pure constitutive explanation. Salmon gives the explanation of "the presence of a worked bone that is thirty thousand years old in an Alaskan archaeological site" as an example of a pure case of etiological explanation, noting that "to explain this fact, it is not essential to look for the causal constituents of the bone" (1984, p. 270).[5]

In general, the new mechanists seem to agree with Salmon that most explanations include both etiological and constitutive aspects; however, whereas Salmon's account emphasizes etiological explanations, the new mechanist philosophy emphasizes constitutive ones. Indeed, Craver explicitly distinguishes his project from Salmon's in exactly this way, stating, "The variety of explanation that I am interested in is constitutive (or componential) causal-mechanical explanation: the explanation of a phenomenon, such as the opening of a Ca^{2+} channel, by the organization of component entities and activities" (2007, p. 8). Similarly, Bechtel acknowledges that "mechanistic explanations are inherently reductionistic insofar as they require specifying the parts of a mechanism and the operations the parts perform" (2011, p. 538). Thus, Darden's example of the mechanism for the segregation of genes seems to be the exception rather than the rule, and Glennan distinguishes ephemeral mechanisms from his primary account of *systems* mechanisms, which *do* involve the decomposition of a system into parts (Glennan 2010a, p. 258).[6]

[5] He also states, "Microphysics is invoked to ascertain the age of the bone, but not explain its presence in the site where it was discovered" (1984, p. 268).

[6] Illari and Williamson (2010) also seem to understand MDC mechanisms as being decompositional. Kuorikoski (2009) usefully distinguishes between mechanisms that involve decomposition and those that do not; he agrees with Skipper and Millstein (2005) that natural selection falls into the latter category. (Thanks to Till Gruene-Yanoff for the pointer to the paper by Kuorikoski).

So, to return to the point at hand, recall my claim that, contra Glennan, there seems to be at least prima facie reason to think that populations of water bugs as a whole often do interact with other entities as a whole. If it were the case that a population could interact with other entities as a whole to produce changes in the very same population, then a pure "at a level" etiological explanation would better illuminate this phenomenon than a constitutive explanation. Thus, Salmon's account, which emphasizes etiological explanations over decompositional ones – an account that Glennan accepts as providing an account of causal production, as mentioned earlier – is a more promising strategy for characterizing natural selection than the new mechanist accounts, which emphasize decompositional explanation over etiological explanation.[7]

To that end, let me briefly review Salmon's views. Salmon's (1984) account[8] describes both *causal propagation* and *causal production*. Salmon suggests that a baseball at rest or in motion is a causal process because it is capable of transmitting (or *propagating*) a mark through time without further interactions. For example, if one makes a scuff on a baseball, the scuff simply persists on the baseball; the baseball, with its mark, propagates through time. On the other hand, changes in causal processes are *produced* by causal interactions, that is, intersections of processes where changes in the characteristics of the processes occur at and persist beyond the space-time point of intersection. For example, the interaction of a moving baseball (a causal process) and a window (another causal process) can produce a change in both the window and the baseball, namely, the breaking of the widow and a change in the trajectory of the baseball. Note that there is no decomposition here; neither the baseball nor the window needs to be broken down into parts in order to explain the interaction between the two causal processes or the production of change. Indeed, the mass of the entire ball is one factor (aside from velocity, wind resistance, etc.) in the window's breaking exactly the way it did.[9]

[7] Skipper and Millstein (2005) offer additional reasons for thinking that the new mechanistic philosophy does not, in its current form, adequately characterize natural selection. I have focused on the issue of decomposition here in order to address the decompositional assumption behind Glennan's claim that population-level properties do not produce change because the population is not a part of the mechanism that produces changes in genotype and phenotype frequencies. I thus seek to highlight the way in which Salmon's account can provide a non-decompositional picture of causal production in natural selection.

[8] I focus on Salmon's Mark Transmission account rather than his later Conserved Quantity account because I believe that it is more broadly applicable to causation outside the domain of physics. Indeed, Salmon explicitly states that his 1984 account of scientific explanation is intended to cover many different disciplines, such as the behavioral sciences, the physical sciences, and the biomedical sciences (1984, p. 267).

[9] Similarly, Salmon notes that when two moving pool balls intersect in space-time, energy and momentum are transferred, altering the states of motion of both balls; thus, the intersection is a causal interaction in which the change in each process can be said to be produced by the other process (1984, pp. 169–170).

Although Salmon distinguishes between causal propagation and causal production, it seems to me that causal propagation can be construed as a type of causal production, or, at least, it can be construed as causal production in Glennan's sense. Recall that, according to Glennan, causal production is (1) a relation between events, (2) local, and (3) transitive. Propagation very clearly satisfies all three of these criteria. The events in question are the ball at one point in space-time and the ball at a subsequent point in space-time; these events are contiguous in time and space and, given a third event in space-time, transitive. I will refer to my claim that causal propagation is a type of causal production later in the chapter.

Now suppose that, like a baseball, a *population* were capable of transmitting a mark; it would then be considered a causal process on Salmon's account, capable of propagating causal influence through space and time. If so, the population could interact with other causal processes, producing a change in the characteristics of those processes at the same time that the other processes produced a change in the characteristics of the population. Then it would seem as though a population could be causally productive of its own changes without citing the activities of the organisms that compose it. But for this to be the case, a population would need to be an object (categorically, and not just in certain situations), so let us turn to that question.

4.2 Populations as Individuals

Elsewhere (Millstein 2009, 2010), I argue that populations are individuals ("objects"), using the Ghiselin-Hull individuality thesis as my inspiration (Ghiselin 1974, 1997; Hull 1976, 1978, 1980). Briefly, my argument is that populations are composed of individual organisms, just as organisms are composed of individual cells; a population is a particular thing – not a class, since it exists in space and time, and not merely a set, since it is integrated via the survival and reproductive interactions of its constituent members with members having a shared fate (albeit less so than organisms); a population has a beginning in time (e.g., migration of organisms away from a population) and an ending in time (e.g., death of the last organism in a population); a population does change over time, but so do organisms; and a population is continuous in time via the causal interactions that occur over time.[10]

[10] Here one might worry about circularity if individuals ("objects") are characterized in terms of interactions, if causal processes are objects persisting and changing through space-time, and if interactions are intersections of causal processes. However, Salmon (1994) clarifies that interactions are not to be defined in terms of causal processes, only in terms of processes more generally, where "[a] process is something that displays consistency of characteristics" (1994, p. 299). Causal processes are then characterized by their ability to transmit marks, where a mark is a type of interaction – "an alteration to a characteristic that occurs in a single local intersection" (Salmon 1994, p. 299). An object persisting or changing through space-time is one *example* of a causal process; however, a carrier wave is another.

The main aspect that Glennan seems to miss here is the extent to which populations are integrated. Recall his claim that "when a fish kills a water bug, it doesn't kill the whole population of water bugs. The life of one water bug is more or less independent of another" (2009, p. 333). But it is not true that the life of one water bug is more or less independent of another. If a fish kills a water bug, then there are more resources (e.g., food, mates) available for the other water bugs. Conversely, a water bug who is adept at obtaining food and mates affects other bugs because those resources are no longer available to them. Indeed, on my view, populations are characterized by their survival and reproductive interactions, with the boundaries of the population as the largest grouping where the rates of interaction are much higher within the grouping than outside. Thus, it seems prima facie as though populations *can* be causally productive in the process of natural selection.

4.3 Potential Worries

But further worries remain. Glennan implies that for populations to be causally productive, they would need to (1) be localized in space and time, (2) have a stable structure, (3) engage in activities as a unified entity in particular times and places, (4) be individuals in the natural selection process, and (5) have parts that share a common fate. I will take up each of these criteria one at a time and show that populations do, contra Glennan, in fact meet them.

With respect to localization in space and time, Glennan worries that in the water bug scenario, "the population as a whole is spread out," which is certainly the case. But there are spaces between the cells that compose an organism, and yet, there is no difficulty conceiving organisms as individuals ("objects"). So, the issue is not space per se; rather, the issue is whether the parts are close enough in space and time so that they can be interacted with as a whole. In Glennan's natural selection example, the predator fish is able to form a stereotypic image of the water bug with the more common color, suggesting that the predator fish is able to perceive the population (or at least a significant percentage of it) *as a whole*. Thus, the population *is* sufficiently localized in space and time to engage in causal production.

The second worry is that populations are not sufficiently stable in the face of interventions to interact as a whole, and it is true that populations are not entirely stable. Even without changes in the environment ("interventions"), organisms may be born (increasing the size of the population) or die (decreasing the size of the populations). Immigration or emigration may also change the size of the population. However, consider fire (a type of "intervention") – a process that would destabilize many otherwise stable entities. Even if many of the organisms of a population were to die in a large fire, the population would generally still retain many of the characteristics that it had before the fire: it would be composed of members of the same species that it was composed of before the fire, some of the same organisms would remain, and some of the genetic and trait variations would remain. Thus, populations seem sufficiently stable to engage in causal production.

The third worry is that populations do not seem to engage in activities as a unified entity in particular places and times. Here, it is not entirely clear what counts as an activity or whether Glennan means to fully take on the MDC notion of activity (which is itself not entirely clear). However, here are some candidate activities that populations can engage in as a whole: invading (as discussed above), changing other populations (e.g., as with predator/prey interactions), splitting, going extinct, speciating, and changing their environments in a way that facilitates colonization by populations of other species. Indeed, if it turns out that that these do not count as activities, so much the worse for the requirement that entities engage in activities. They all involve *interactions* (the term used in Glennan's own account of mechanisms) between populations and other entities, including interactions between populations and entities in the populations' environments.

Glennan does acknowledge that populations sometimes act as individuals, for example, in migration processes. However, he says, "With respect to selection processes, the question of whether or not populations or sub-populations should be treated as individual entities depends upon whether or not group selection is at work" (Glennan 2009, p. 333). More generally, "The question of whether they are individuals only makes sense in the context of analyzing a particular causal process" (Glennan 2009, p. 333). So, this raises a fourth worry, whether populations are individuals in the natural selection process specifically.

However, it seems to me that the population *is* acting as an individual with respect to the selection example that Glennan describes, even in the absence of group selection. Again, recall that the fish form an image associated with the more common color. This in itself is evidence that there is an interaction between the fish and the population as a whole – the fish forms an impression of the population as a whole, and the image is a result of the interaction. Of course, when a predator fish kills a water bug, there is an interaction between an individual predator and an individual bug. But that single interaction does not constitute a natural selection process, just as the interaction of your fingers with a keyboard does not constitute the creation of a document; that involves your interaction with the whole computer. Or, to invoke an analogy for selection processes more generally, a single particle of flour falling through the hole of a sifter does not constitute sifting. One sifts not a single particle of flour, but rather a "population" of flour particles, with particles jostling against each other, some falling through and some remaining in the sifter. Similarly, selection occurs with respect to the whole population. Types are only selectively favored or disfavored as compared to other types in the population; a type that might appear reproductively successful when considered individually is actually unsuccessful in the selection process if other types outreproduce it (Millstein 2006). Thus, for selection in general, the population acts as an individual.

Finally, there is the worry that populations do not have parts that share a common fate. However, the fact that the organisms (the "parts") of a population are engaging in survival and reproductive interactions implies that they do have a shared fate, at least to some extent. For example, consider a new advantageous variation introduced into a population. If there is interbreeding among the organisms (one kind of reproductive interaction), then that variation may spread

in the population, enhancing the survival of the population as a whole. Indeed, there are many kinds of interactions among the members of the population.[11] Survival interactions include direct physical combat; competition for limited food, sunlight, or shelter resources; and cooperation, whereas reproductive interactions include mating successfully or unsuccessfully and offspring rearing. Lots of interactions imply that the organisms will share a common fate to a high degree.

To summarize, I have argued that populations, to a sufficient extent, are categorically individuals (objects) and are localized in space and time, that they do have a stable structure, that they engage in activities as a unified entity, and that the members of a population share a common fate. Thus, populations are not excluded from being causally productive on that basis. But to make the positive case for populations as causally productive, I return to Salmon's account of causal propagation, causal production, and the baseball example, which I use as an analogy.

5 Populations Can Be Causally Productive

First, like a baseball, a population is capable of transmitting a mark. For example, if an organism in the population is born or killed, that "mark" persists in future states of the population. However, Michael Strevens (personal communication) raises the worry that if an organism disappearing from the population counts as a mark, then Salmon's criterion will collapse. According to Strevens, Salmon wants to say, for example, that a shadow traveling across a wall is not a causal process because "marks" made on the shadow at one point (e.g., by a blemish on the wall) do not persist to the next point – but the effect on a population of killing a member seems very much like that (at one moment there, at the next moment not). Here I would respond that, on my account, an organism is a member of a population in virtue of the fact that it is interacting with other members of the population. So, if a new organism is born, it will affect other organisms: eat their food, offer them some food, mate with them, refuse to mate with them, etc. The population is changed because of that new organism. So, when that organism later dies, the rest of population is similarly affected – perhaps a small amount, but an effect nonetheless. And since most organisms are more than just ephemeral shadows (let us suppose most of them live more or less the average for the species), I think their appearance and disappearance is different than the appearance and disappearance of a shadow. The organisms persist, and thus, the mark on the population persists as well. That being said, there are probably more obvious sorts of marks, such as a disease that quickly spreads through a population, and, of course, all that really needs to be

[11] The interactions within (or among) the members of a population are to be distinguished from the interactions between the population as a whole and other entities. It is the occurrence of the former interactions that binds the population together as a whole and thus makes possible the latter kinds of interactions.

shown for a population to be a causal process on Salmon's account is that a population is *capable* of transmitting a mark. My point in choosing birth and death as examples of marks is that mark transmission is not only possible for a population, it is commonplace.

Second, like a baseball at rest or in motion, the state of the population (the entity) at one point in space-time can *propagate* its influence to another point in space-time simply by persisting or even while changing (e.g., moving). The genotype and phenotype frequencies of a population at one point in time probabilistically[12] propagate the genotype and phenotype frequencies to future points in time. This propagation is reflected in transition probability models, equations that describe the probability of various possible future states, given the current state of a population. Future states of the population are partially the result of, and are constrained by, present states. As I suggested above, this propagation itself is a type of causal production, albeit different than the type of causal production that occurs as the result of an interaction.

Third, like a baseball that hits a window, the population can *produce* changes in other causal processes through causal interactions and be changed in turn. As Skipper and Millstein (2005, p. 345) suggest, "To capture natural selection as a mechanism, an account of productive continuity is required that captures the ways in which relevant property differences among a population of entities entering into causal interactions with their environment is productive of change in that population." To return to Glennan's example, recall that each predator fish is forming a stereotypic searching image representing the most common water bug color in the population; this is an interaction between the fish and the water bug population. Thus, we can say that a population of water bugs, with dark forms rarer, repeatedly interacts with predator fish to probabilistically produce relative increases in the darker form as a result of preferential predation (discriminate sampling) of the lighter forms. In this way, natural selection can, contra Glennan, exhibit causal production at the population level.

Or consider one of the cases discussed in Skipper and Millstein (2005) where frequency-dependent selection is not involved. Suppose there exists a population of finches that vary in their beak length, a heritable trait, with the varying beak lengths conferring variable abilities to obtain seeds for food. The population of finches repeatedly interacts with the seeds in the environment, so that some finches are favored over others based on the differences among the finches, producing future changes in distributions of types in the population. In other words, the environment (in the form of seeds) discriminates among the members of the finch population; this interaction between population and environment produces changes in both the population and the environment (analogous to an interaction between flour and a sifter). Again, it is of course true that a particular finch can also interact with a particular seed, but that interaction neither constitutes selection nor prevents an

[12] Salmon intends his account to include probabilistic processes; see, for example, his 1984 work, p. 268.

interaction from occurring at the population level. This case illustrates how even in non-frequency-dependent situations, natural selection exhibits causal production at the population level.

These three points taken together clarify the way in which populations can be seen as causally productive via Salmon's Mark Transmission account. However, I must make a few caveats. I am not endorsing Salmon's account over other accounts of causation or mechanisms; other accounts may be needed to supplement Salmon's account or may handle other sorts of cases better.[13] Moreover, I do not think this discussion of Salmon's account captures everything there is to be said about natural selection as a population-level causal process (or, to be consistent with Salmon's terminology, perhaps I should say "natural selection as a population-level *causal nexus*," since there are many interactions between populations and their environments and populations and other organisms); for example, I have not said anything about the way in which natural selection can be distinguished from other, similar causal processes, such as sexual selection and artificial selection. But I do think that Salmon's views on causal propagation and causal production can capture some important aspects of the role of populations in natural selection. His views help elucidate the ways in which populations propagate their influence through space and time as well as the ways in which populations' interactions with various other entities in their environment produce changes in those populations.

6 Conclusions

My main goal in this chapter has been to respond to Glennan; Glennan argues that entities like populations can only give rise to causally relevant causes in the process of natural selection, but as I have sought to show, populations can be causally

[13] One worry that has been raised by a number of recent authors, including Glennan (Glennan 2009; see also Hitchcock 1995 and Craver 2007), is that Salmon's account fails to pinpoint which of the causal processes that produce an effect are explanatorily relevant. In one version of an example which purports to illustrate the problem, Ms. Slims chalks her cue stick with blue chalk and deftly hits the cue ball, which hits the eight ball, which proceeds to the corner pocket. The claim seems to be that, while the blue "mark" has been transmitted (perhaps even to the eight ball), it is not explanatorily relevant to the effect. However, I think we need to be clear on what the effect is; if we are talking about a token chain of events (and not a type of chain of events), then the effect that occurred is that an eight ball with a blue mark dropped into a corner pocket. And the blue mark is explanatorily relevant to that token event, just as the momentum of the cue ball is. We still might be worried that Salmon wanted his account to be able to give an explanation for the event type "ball in the corner pocket" and that the blue mark is not relevant to that. Here, I think three possible responses are open. One is that explanatory relevance and causal relevance come apart; the blue mark is always causally relevant, but it simply is not explanatorily relevant to the event type. Second is to insist that in explaining why an eight ball with a blue mark has gone into the corner pocket, we have already explained why the eight ball has gone into the corner pocket. Third is to give up on using Salmon's account to explain event types and only use it to explain event tokens. (Thanks to Christopher Hitchcock for helpful discussion).

productive, too, both through causal propagation and causal interactions. I was initially motivated to respond to Glennan because it seemed to me that, if correct, his claims would imply that the merely causally relevant causes occurring at the population level were somehow "lesser" than the more robust causally productive causes at the level of individual organisms. These thoughts probably have more to do with my views about causation than Glennan's, although I am apparently not alone in this way of thinking; as Jaegwon Kim suggests, "causal production, which respects the locality/contiguity condition," involves "*real connectedness* between cause and effect" (2007, p. 236; emphasis in original). Furthermore, Glennan claims that full understanding of the causal basis of an event requires *both* the causally productive causes and the causally relevant causes even though he believes that at the population level of natural selection, there can be causal relevance without causal production. This seems to leave bare causal relevance at the population level a bit free-floating and weird. Finally, although in this chapter I have not sought to question the claim that there are two kinds of causes, I find it somewhat troubling. For all of these reasons, it seemed to me that he was mounting a serious challenge to my claim that natural selection is a population-level causal process (Millstein 2006): that those population-level causes were "lesser" or "free floating and weird" or part of a distinction that was not fully coherent and thus perhaps ephemeral. So, responding to Glennan here is, in part, a defense of my earlier work.

However, I hope to have made some other, more general points along the way. One is that while I find the new mechanists' approach appealing for many areas of biology (such as molecular biology and neuroscience), I do not think it illuminates all cases. This echoes a claim of Skipper and Millstein (2005), but here I go beyond that negative claim to show how Salmon's Mark Transmission account can be more helpful in understanding other sorts of biological phenomena, such as natural selection. Salmon eventually abandoned his Mark Transmission account because he felt it relied too much on counterfactuals; however, for people like me who do not find counterfactuals ontologically objectionable (and anyone who defends a causal dependence view of causality cannot find counterfactuals ontologically objectionable), there is much insight to be gained by analyzing cases in terms of Salmon's account. In part, this is because (as I argued above) phenomena such as natural selection are better suited to non-decompositional, etiological accounts, rather than the constitutive decompositional accounts that the new mechanists emphasize. Sometimes, all we need is to cite causation "at" a level. However, I also think that concepts such as "causal processes," "causal propagation," and "causal interaction" are rich and powerful tools. I recommend Salmon's Mark Transmission account as an alternative to the new mechanists' approach – again, not as a replacement but as a supplement. I expect that other areas of biology and science more generally might be fruitfully examined through the lens of Mark Transmission. Whether Salmon's account should itself be considered a type of mechanist approach is a matter for another time, and I do not think anything I have said here turns on that question.

Finally, I think it is important that we understand what sort of entities can enter into causal relations and in what ways. I think we have certain human-centered

biases about what entities count as individuals, and these biases can lead us to mistaken conclusions about causality. If populations can be causally productive, perhaps other, similar entities can as well: communities, ecosystems, etc. Organisms are not a privileged level of organization.

Acknowledgements Thanks to Carl Craver, Lindley Darden, and Stuart Glennan for much relevant and productive discussion about causation and mechanisms. Thanks also to the Griesemer/Millstein Lab, my Winter 2011 Philosophy of Science seminar, and attendees of the Taiwan Conference on the Philosophy of Biology and Economics for helpful comments and questions, and to Joyce Havstad and Michael Strevens for helpful comments on my draft. Finally, thanks to Carl Craver for an excellent set of referee comments.

References

Bechtel, William. 2011. Mechanism and biological explanation. *Philosophy of Science* 78: 533–557.

Bechtel, William, and Adele Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Cartwright, Nancy. 2004. Causation: One word, many things. *Philosophy and Phenomenological Research* 71: 805–819.

Craver, Carl F. 2001. Role functions, mechanisms, and hierarchy. *Philosophy of Science* 68: 53–74.

Craver, Carl F. 2007. *Explaining the brain*. Oxford: Oxford University Press.

Craver, Carl F., and William Bechtel. 2007. Top-down causation without top-down causes. *Biology and Philosophy* 22: 547–563.

Darden, Lindley. 1991. *Theory change in science: Strategies from Mendelian genetics*. Oxford: Oxford University Press.

Darden, Lindley. 2005. Relations among fields: Mendelian, cytological and molecular mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2): 349–371.

Dowe, Phil. 2000. *Physical causation*. Cambridge: Cambridge University Press.

Ghiselin, Michael T. 1974. A radical solution to the species problem. *Systematic Zoology* 23(4): 536–544.

Ghiselin, Michael T. 1997. *Metaphysics and the origin of species*. Albany: SUNY Press.

Glennan, Stuart. 1996. Mechanisms and the nature of causation. *Erkenntnis* 44: 49–71.

Glennan, Stuart. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69: S342–S353.

Glennan, Stuart. 2009. Productivity, relevance, and natural selection. *Biology and Philosophy* 24: 325–339.

Glennan, Stuart. 2010a. Ephemeral mechanisms and historical explanation. *Erkenntnis* 72: 251–266.

Glennan, Stuart. 2010b. Mechanisms, causes, and the layered model of the world. *Philosophy and Phenomenological Research* LXXXI(2): 362–381.

Hall, Ned. 2004. Two concepts of causation. In *Causation and counterfactuals*, ed. John Collins, Ned Hall, and L.A. Paul, 225–276. Cambridge, MA: The MIT Press.

Hitchcock, Christopher Read. 1995. Salmon on explanatory relevance. *Philosophy of Science* 62 (2): 304–320.

Hull, David L. 1976. Are species really individuals? *Systematic Zoology* 25: 174–191.

Hull, David L. 1978. A matter of individuality. *Philosophy of Science* 45: 335–360.

Hull, David L. 1980. Individuality and selection. *Annual Review of Ecology and Systematics* 11: 311–332.

Illari, Phyllis McKay, and Jon Williamson. 2010. Function and organization: Comparing the mechanisms of protein synthesis and natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences* 41: 279–291.

Kim, Jaegwon. 2007. Causation and mental causation. In *Contemporary debates in philosophy of mind*, ed. Brian McLaughlin and Jonathan Cohen, 227–242. New York: Wiley-Blackwell.

Kuorikoski, Jaakko. 2009. Two concepts of mechanism: Componential causal system and abstract form of interaction. *International Studies in the Philosophy of Science* 23(2): 143–160.

Macan, T.T. 1976. A twenty-one year study of the water-bugs in a moorland fishpond. *Journal of Animal Ecology* 45(3): 913–922.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67: 1–25.

Millstein, Roberta L. 2006. Natural selection as a population-level causal process. *The British Journal for the Philosophy of Science* 57(4): 627–653.

Millstein, Roberta L. 2009. Populations as individuals. *Biological Theory* 4(3).

Millstein, Roberta L. 2010. The concepts of population and metapopulation in evolutionary biology and ecology. In *Evolution since Darwin: The first 150 years*, ed. M.A. Bell, D.J. Futuyma, W.F. Eanes, and J.S. Levinton, 61–86. Sunderland: Sinauer.

Salmon, Wesley. 1984. *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, Wesley. 1994. Causality without counterfactuals. *Philosophy of Science* 61: 297–312.

Skipper, Robert A., and Roberta L. Millstein. 2005. Thinking about evolutionary mechanisms: Natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2): 327–347.

Chapter 9
Is Natural Selection a Population-Level Causal Process?

Rong-Lin Wang

Abstract Recent discussions of natural selection focus on two questions: first, is natural selection a causal process or is it a statistical consequence of lower-level events? And second, is natural selection at the population level or at the level of individuals? Bouchard and Rosenberg (Br J Philos Sci, 55:693–712, 2004) argue that natural selection is causal and at the level of individuals, as opposed to Matthen and Ariew (J Philos, 99:55–83, 2002) and Walsh et al. (Philos Sci, 69:452–473, 2002), who argue that natural selection is at the population level and purely statistical. In addition to these two polar extreme positions, Millstein (Br J Philos Sci, 57:627–653, 2006) tries to steer a middle course by arguing that natural selection is a population-level causal process. I will make three points in this chapter: first, Millstein's account of natural selection is incomplete, in the sense that nowhere in her account can one find a place for cases of natural selection-of. Second, we should prefer Brandon's account of natural selection and drift over Millstein's, on the grounds that her account fails to meet a plausible requirement that Brandon's account succeeds in meeting: namely, whenever natural selection and drift operate together, a change in the strength of natural selection implies an inverse change in the strength of drift, and vice versa. Third, the prospects of the view that natural selection is a population-level causal process depend on a satisfactory solution to both the epiphenomenon and the overdetermination problems. With the help of an analogy, I will show how the two problems can be dealt with.

Recent discussions of natural selection focus on two questions: first, is natural selection a causal process or is it a statistical consequence of lower-level events? And second, is natural selection at the population level or at the level of

R.-L. Wang (✉)
Department of Philosophy, National Taiwan University, 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan
e-mail: rlwang@ntu.edu.tw

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_9,

individuals? Bouchard and Rosenberg (2004) argue that natural selection is causal and at the level of individuals, as opposed to Matthen and Ariew (2002) and Walsh et al. (2002), who argue that natural selection is at the population level and purely statistical.

In addition to these two polar extreme positions, Millstein (2006) tries to steer a middle course by arguing that natural selection is a population-level causal process. On this view, (1) natural selection can be a cause of evolution, namely, it is able to make a change in the frequency of traits among a population from one generation to the next; (2) natural selection is by nature comparative: whenever it acts as a cause of evolution, it impinges on comparative and thus population-level properties (e.g., variation in fitness, frequency of traits) rather than on individual-level properties (e.g., fitness, traits). According to Millstein (2002, 2005, 2006), who endorses a modified version of Beatty's (1984) account, "natural selection should be characterized as a discriminate sampling process whereby physical differences between organisms are causally relevant to differences in reproductive success. Drift, by contrast, is an indiscriminate sampling process whereby physical differences between organisms are causally irrelevant to differences in reproductive success" (2006, p. 640). Such a sampling process, discriminate or not, operates at the population level.

I will make three points in this chapter: first, Millstein's account of natural selection is incomplete, in the sense that nowhere in her account can one find a place for cases of natural selection-*of*. Second, we should prefer Brandon's account of natural selection and drift over Millstein's, on the grounds that her account fails to meet a plausible requirement that Brandon's account succeeds in meeting: namely, whenever natural selection and drift operate together, a change in the strength of natural selection implies an inverse change in the strength of drift, and vice versa. Third, the prospects of the view that natural selection is a population-level causal process depend on a satisfactory solution to both the epiphenomenon and the overdetermination problems. With the help of an analogy, I will show how the two problems can be dealt with.

1 Why Is Millstein's Account of Natural Selection Incomplete?

Sober (1984) draws a contrast between selection-*of* and selection-*for*. To see why, consider the following two propositions:

(1) There is selection-*of* trait T in population p if and only if T is fitter than not-T in p.
(2) There is selection-*for* trait T in population p if and only if T is fitter than not-T in p.

According to Sober, (1) is true but (2) is false. That is how selection-of differs from selection-for. Here is a case of selection-of:

> Consider a population in which there is selection for being green; this selection pressure exists because being green camouflages organisms in the green environment they occupy, thus protecting them from predators. Suppose further that there is no selection for being small—body size is selectively irrelevant. And now imagine that all and only the green organisms in the population are small. In this situation, the green organisms are selected, which means that the small ones are too. However, though there is selection for being green, there is no selection for being small. Selection-of is the concept that is tightly connected to variation in fitness; if there is selection of green (small) organisms, then they are on average fitter than those that are not. But the fact that small organisms are fitter than organisms that are not small does not entail that there is selection for being small. (Shapiro and Sober 2007)

It is noteworthy that selection-of is a *discriminate* sampling process in the sense that such a process of selection is not random. In fact, selection-of is a sampling process which, based on differences in fitness between organisms, favors fitter organisms. And that is how small organisms are favored and selected. On the other hand, it is also noteworthy that, in the course of selection-of process, differences in fitness are *causally irrelevant* to differences in survival and reproductive success among organisms. The fact that small organisms are fitter is causally irrelevant to that fact that small organisms have a greater success in survival and reproduction. Small organisms are selected accidentally.

Now, let's turn to selection-for. It is beyond question that selection-for is a discriminate sampling process. And such a sampling process, to be sure, is no less discriminate than selection-of. Still, selection-for is distinguishable from selection-of, and that is why Sober draws a contrast between them. In the case of selection-for, as opposed to selection-of, differences in fitness among organisms are *causally relevant* to differences in survival and reproductive success. Recall that Millstein characterizes natural selection as "a discriminate sampling process whereby physical differences between organisms are causally relevant to differences in reproductive success." Such a characterization, as one can see now, fits well and only well with selection-for. As to selection-of, it is missing in Millstein's account of natural selection.

Consider Brandon's (2005) account of natural selection. Brandon takes natural selection to be a discriminate sampling process where discriminate means unequal probability of being chosen in the course of sampling process. So, without any difference in fitness among a population of organisms, there would be no discriminate sampling process. Difference in fitness among organisms is a sine qua non condition for natural selection. In this sense, Brandon's account has the advantage of putting natural selection-of into its place: natural selection is a discriminate sampling process including both selection-for and selection-of processes. By contrast, in Millstein's account, there is a conflation between natural selection-of with drift if (1) between the discriminate and the indiscriminate sampling processes that she characterizes, there is no third option and if (2) by drift she means each and every sampling process whereby physical differences between organisms are causally irrelevant to differences in reproductive success.

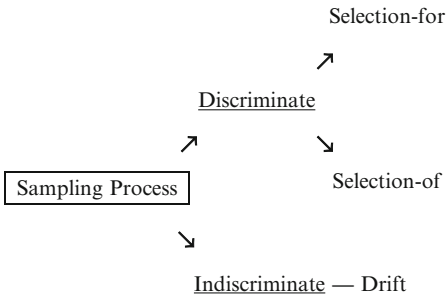


Fig. 9.1 Distinction between selection-for, selection-of, and drift in terms of discriminate/indiscriminate sampling processes

Although Millstein succeeds in offering a right characterization of natural selection-*for*, she does not offer a right characterization of natural selection tout court. In contrast, Brandon's account of natural selection is right, but he fails to make a distinction between selection-for and selection-of. In this regard, Millstein and Brandon are half right and half wrong with respect to the characterization of natural selection. In my view, a more complete picture would be like this (Fig. 9.1):

And a more satisfactory distinction would be as follows:

Natural selection-for: a *discriminate* sampling process whereby physical differences between organisms are *causally relevant* to differences in reproductive success

Natural selection-of: a *discriminate* sampling process whereby physical differences between organisms are *causally irrelevant* to differences in reproductive success

Drift: an *indiscriminate* sampling process whereby physical differences between organisms are *causally irrelevant* to differences in reproductive success

One might wonder whether I use the term "discriminate" in an equivocal way: on the one hand, "discriminate" means sampling processes in which differences in reproductive success are connected to differences in fitness (i.e., discriminate between fitter and less fit traits). On the other hand, "discriminate" means sampling processes whereby differences in reproductive success are causally relevant to physical differences (i.e., discriminate between causally more and causally less accountable traits). Indeed, given my intention to use the term in a broader sense so as to cover both selection-for and selection-of, it is not surprising that the problem of equivocality arises.

To deal with the problem, let me draw a distinction between the two senses of "discriminate" and call them, respectively, *F*-discriminate and *C*-discriminate. The term "*F*-discriminate" is intended to refer to discrimination between fitter and less fit traits. In *F*-discriminate sampling processes, physical differences among organisms are still *statistically*, though *not causally*, relevant to differences in reproductive success. As for the term "*C*-discriminate," it refers to discrimination between *causally* more and *causally* less accountable traits. In *C*-discriminate sampling processes, physical differences among organisms are *causally* relevant to differences in reproductive success. It turns out that selection-of is an *F*-discriminate sampling process, whereas selection-for is a *C*-discriminate sampling process.

Generally speaking, causal relevance, as is well known, implies statistical relevance. That means in *C*-discriminate sampling processes, physical differences among organisms are *statistically* relevant to differences in reproductive success. So *F*-discriminate and *C*-discriminate sampling processes turn out to have something in common, namely, statistical relevance. It is statistical relevance as common core that permits the term "discriminate" to have a meaning broad enough to cover both selection-for and selection-of. Now, the term "discriminate," in its broader sense, means sampling processes in which physical differences among organisms are *statistically* relevant to differences in reproductive success.

So a solution to the problem of equivocality is available, and it helps to secure the characterization of selection-of as a kind of discriminate sampling process, in contrast to Millstein's account. Despite the solution, I suspect that what is really at issue concerns not whether I use the term "discriminate" equivocally; the real issue concerns instead why natural selection, as Millstein proposes, should be confined to *C*-discriminate sampling processes alone. Indeed, Millstein's account of natural selection is causalist in its entirety, which definitely bars *F*-discriminate sampling processes, in general, and selection-of, in particular, from counting as selection processes. It is interesting to note that Millstein never characterizes natural selection in terms of fitness. Although Millstein (personal communication) does not doubt that one can have a fitness-based account of natural selection, she avoids using the term "fitness" because it is a controversial term, and she does not need it to make the points she wants to make. In Millstein's view, a causalist account of natural selection remains defensible without resort to the contentious term "fitness." I surely agree with her on this point. Nonetheless, it seems appropriate to separate two questions apart: what is an adequate causalist account of natural selection? And what is an adequate account of natural selection tout court? Indeed, though I agree that Millstein succeeds in offering an adequate causalist account, I have been arguing contra Millstein that the idea of natural selection should be broadened enough to include *F*-discriminate sampling processes. Given that the argument I offered above is based on the controversial term "fitness," its strength accordingly is limited if it is to be addressed to a causalist like Millstein. So let me try not to argue against her in terms of fitness.

Consider again the case of selection-of. Selection-of is a sampling process that biologists often name "hitchhiking." Such a name suggests that selection-of cannot occur all by itself: whenever there is selection-of, there is selection-for. Let's call such a property of selection-of *hitchhikliness*. Needless to say, the property of hitchhikliness is essential to selection-of. Recall that on Millstein's causalist account of natural selection, there is no way to count selection-of as a discriminate sampling process. Thus, whenever selection-of and selection-for are both present, granting that selection-for is a discriminate sampling process, there must be two, rather than one, sampling processes. In short, Millstein's causalist account implies that it is impossible for selection-of to be *numerically* identical to selection-for. However, such an implication is problematic, for it surely is possible that a selection-of sampling process is numerically identical to a selection-for sampling process. I would say the property of hitchhikliness, which is essential to selection-of, provides evidence in support of this claim: selection-of and selection-for on

which it hitchhikes *actually* are numerically identical. Namely, despite the appearance, they are, ontologically speaking, one and the same sampling process. It is noteworthy that in order to show the implication of Millstein's account is problematic, such a strong claim is *not* necessary. A weaker claim would suffice: when selection-of and selection-for are both present, it is *possible* that only one single sampling process is occurring. But is it really possible?

Recall the example of selection-of offered in Shapiro and Sober (2007). Ontologically speaking, selection for being green and selection of being small are not two separate sampling processes. The distinction between them is made in the human mind, not in the world. It is we who separate selection of being small from selection for being green. Nature never separates the two sampling processes. Green organisms and small organisms are selected all at once and in one shot. It is not the case that green organisms are first selected, and then it's the turn of small organisms to be selected. Being numerically identical, selection for being green and selection of being small are one and the same sampling process. Since selection-of and selection-for are, arguably, one single sampling process, an adequate account of natural selection should be broad enough to include both selection-for and selection-of. The problem with Millstein's account is that it defines discriminate sampling process (natural selection) too narrowly to include selection-of: not only is it impossible for selection-of and selection-for to be one single sampling process, it is also impossible that natural selection and drift, when operating together, are one and the same sampling process. As is clear now, if a sampling process is discriminate, then it surely is not indiscriminate. Thus, if it counts as natural selection, then it cannot count as drift, and vice versa. In short, natural selection and drift, according to Millstein, are mutually exclusive. In my view, Millstein should leave open the possibility that natural selection and drift coact as one single sampling process. This includes two things: first, it is possible that natural selection and drift act together. Second, it is possible that when they act together, only one single sampling process is occurring.

Let's consider the first requirement. Indeed, the coexistence of natural selection and drift, far from being unusual, makes better sense of biologists' practice. When biologists debate whether natural selection or drift predominates in an evolutionary event, they are presupposing that natural selection and drift may act together. Otherwise, their debate, if not pointless, would be highly misleading. Note that although Millstein characterizes natural selection and drift as mutually exclusive, it does not follow that they cannot act together. To be sure, a sampling process, if it counts as natural selection, will not be drift, and if it counts as drift, it will not be natural selection. However, it is possible that two sampling processes act together, one of which counts as natural selection and the other as drift. Hence, it is possible that natural selection and drift act together. In this sense, Millstein would have no problem with the first requirement.

As is evident now, I will argue that Millstein should meet the second requirement, namely, when natural selection and drift act together, it is possible that only one single sampling process is occurring. In my view, this is a move that Millstein should make. Although making such a move may not require Millstein to abandon

the idea that natural selection is characterized as a (*C*-)discriminate sampling process, it does require Millstein to revise her characterization of drift so as to make room for its being numerically identical to natural selection when they act jointly. I will begin my argument by pinpointing a problem with Millstein's account of drift. That brings me to a disagreement between Brandon and Millstein with respect to drift.

2 Why Is Brandon's Account of Natural Selection and Drift Preferable?

According to Millstein, natural selection, as a causal process, is probabilistic. That is why she characterizes natural selection in terms of causal *relevance* rather than causal *determination*. When natural selection acts, it is true that physical differences among organisms are causally relevant to differences in reproductive success. However, the physical differences do not causally determine the differences in reproductive success: given the same physical differences, different differences in reproductive success are possible. Among the many possible results, there may well be a distribution of probabilities, and every possible result is not equally probable. The most probable result is surely the most expected when natural selection acts, but the other results, even improbable, remain possible. That is how natural selection, though a causal process, is nonetheless probabilistic.

Recall that Millstein characterizes drift, in terms of causal irrelevance, as an indiscriminate sampling process. When drift occurs, physical differences among organisms are not causally relevant to differences in reproductive success, let alone causally determine them. Thus, drift is no less *probabilistic* than natural selection. Like natural selection, drift has many possible results, among which there may well be a distribution of probabilities, and every possible result may not be equally probable.

Millstein claims, on the one hand, that the distinction in terms of discriminate/indiscriminate sampling process alone suffices to separate, at least conceptually, natural selection from drift. On the other hand, she claims that an improbable result of natural selection might be indistinguishable from a product of drift. At first sight, it appears paradoxical to make the two claims at once. Millstein argues, however, that one should separate process from outcome in order to get things clear: although natural selection as outcome might be indistinguishable from drift as outcome, natural selection as process is conceptually distinguishable from drift as process. And such a distinction can be made without regard to outcomes.

To Millstein's view, Brandon (2010) raises an objection:

> [N]atural selection and drift are co-products of the same process, namely a probabilistic sampling process (Brandon and Carson 1996; Matthen and Ariew 2002; Walsh et al. 2002). Thus, although it is of crucial importance to separate selection and drift, one cannot do so on the basis of process alone (contra Millstein 2002), one must do so on the basis of outcomes. (Brandon 2010)

According to Brandon, random drift is any deviation from expected result due to sampling error.[1] Now, sampling error occurs whenever the sampling process is unrepresentative. Thus, any deviation from the expected result is a proof that the sampling process is unrepresentative; conversely, only deviation can prove that the sampling process is unrepresentative. On this account, deviation and sampling error/unrepresentative sampling are, so to speak, two sides of one coin. And that is why, on Brandon's view, drift as process (viz., sampling error/unrepresentative sampling) cannot be defined independently of drift as outcome (viz., deviation from expected result).[2] For Brandon, deviation and only deviation indicates a sampling process is not representative, which in turn requires that the effective population size be finite.[3]

Not surprisingly, Millstein disagrees with Brandon's account of drift. As mentioned above, the idea of "fitness" on which Brandon heavily relies to make his case is highly controversial, and Millstein does not need it to make the points she wants to make. In my view, one can appreciate Brandon's account without drawing on the notion "fitness." And that is why, as shown in the preceding paragraph, I rephrased his remarks without using "fitness." To parallel Millstein's process-oriented account, I would say what is essential to Brandon's view is that drift is an *unrepresentative* sampling process.

Millstein (2005) points out a second problem with Brandon's view: granting that natural selection is a *probabilistic* sampling process, any result, with or without deviation, is no proof of whether natural selection as process has occurred. And since drift is also a *probabilistic* sampling process, the same can be said of drift as process. So contrary to Brandon's view, outcome is neither sufficient nor necessary to distinguish conceptually natural selection from drift.

Granting that discriminate sampling process is probabilistic and that indiscriminate sampling process is probabilistic as well, I think Millstein is right that process alone suffices to make a conceptual distinction between them. But it does *not* follow that process alone suffices to make a conceptual distinction between natural selection and drift. Unless it is the case that natural selection surely is a discriminate sampling process and that drift surely is an indiscriminate sampling process, Millstein would not be justified in inferring that outcome is neither sufficient nor necessary to distinguish conceptually natural selection from drift. As it turns out, the crucial issue is whether drift surely is an indiscriminate sampling process.

[1] As an anonymous referee points out, Brandon's views on drift may have changed since 2005. Because my paper focuses on Brandon's (2005) argument where he responds directly to Millstein, I will not refer to Brandon's (2006) "The Principle of Drift: Biology's First Law," where he seems to offer a fleshed out alternative view of drift.

[2] Note that Brandon deliberately separates deviation from sampling error in using the expression "due to," which suggests, among others, that he views deviation as outcome and unrepresentative sampling/sampling error as process.

[3] In an infinite population, it is extremely unlikely that deviation arises. Alternatively speaking, it is extremely likely that sampling process is representative.

The idea that drift is equivalent to an indiscriminate sampling process, as shown in Millstein's account, is problematic. According to Millstein, natural selection is a causal process in the sense that physical differences among organisms are causally *relevant* to differences in reproductive success. The strength of natural selection depends accordingly on the degree of causal relevance. By contrast, drift is characterized in terms of causal *irrelevance*. Can the strength of drift depend on the degree of causal irrelevance? Note that on Millstein's account, causal irrelevance, unlike causal relevance, cannot be a matter of degree. Causal irrelevance is a matter of all or nothing, not a matter of more or less. If causal irrelevance were a matter of degree, granting that "highly causally irrelevant" is equivalent to "barely causally relevant" and that "moderately causally irrelevant" to "moderately causally relevant," the clear-cut distinction between natural selection and drift would be blurred. As a result, there might be cases where one and the same sampling process is both natural selection and drift, with a variation in degree of strength. Such cases would threaten to undermine Millstein's efforts to distinguish conceptually natural selection from drift. In short, causal irrelevance, unlike causal relevance, cannot be a matter of degree. Hence, although the strength of natural selection varies with the degree of causal relevance, the strength of drift can vary neither with the degree of causal irrelevance nor with the degree of causal relevance. In other words, causal relevance is a factor of strength in the case of natural selection, but not in the case of drift.

Let's consider a second factor that affects, as Millstein (personal communication) agrees, not only the strength of natural selection but also the strength of drift, namely, the effective population size. Indeed, the effective population size, as a factor of strength, affects any sampling process, *discriminate or not*. Now, is there any connection between the two factors, namely, the size of population and causal relevance? Here is Millstein's favorite example of drift:

> To use Hartl and Clark's example, imagine shellfish that 'produce vast numbers of pelagic larvae that drift about in the sea' (Hartl and Clark 1989, p. 70). Although Hartl and Clark do not elaborate, the image is of virtually identical larvae, subject to the vagaries of tides and predators (i.e., indiscriminate sampling). (Millstein and Skipper 2007)

Note that an indiscriminate sampling process can operate in any population, whatever its size. Pelagic larvae among a large population would drift about in the sea no less than larvae among a small population. Conversely, the size of population has no impact on how indiscriminate a sampling process would be when it operates in the population. The sampling process occurring in a large population of larvae would be as indiscriminate as the sampling process occurring in a small population. Now, if one reads "indiscriminate" as "causal irrelevance," then causal irrelevance has nothing to do with the size of population. A similar inference would lead to the conclusion that causal relevance also has nothing to do with the size of population. Not surprisingly, for any information about the size of population would be unnecessary to tell whether or not physical differences among organisms are causally relevant to differences in reproductive success. It follows that the two

factors of strength, namely, the size of population and causal relevance, are *independent* from each other.

To sum up, on Millstein's view, the strength of natural selection depends on two independent factors, namely, causal relevance and the effective population size. In contrast, the strength of drift depends exclusively on the effective population size. When the degree of causal relevance is higher and the effective population size larger, the strength of natural selection increases. And the larger is the effective population size, the smaller is the strength of drift.

Now, a problem with Millstein's account is that it would lead to a disagreement with most biologists on how to determine the relative strength of natural selection to drift. When natural selection and drift operate together, biologists are interested in telling whether one of them predominates. And their judgment is based on a crucial quantity, namely, *4Ns*, where *N* is the effective population size and *s* is the selection coefficient. If the quantity increases, then natural selection tends to dominate drift; on the contrary, if the quantity decreases, then it is drift that tends to predominate. It is noteworthy that interpreting "s" as "the degree of causal relevance" makes perfect sense of Millstein's view that causal relevance and the effective population size represent indeed two separate factors in determining the strength of natural selection. In addition, since *s* is the *selection* coefficient, it also makes perfect sense of Millstein's view that the effective population size is the only factor that affects the strength of drift. Thus, when both *s* and *N* increase, the crucial quantity *4Ns* increases accordingly, and it turns out that natural selection tends to dominate drift, because the strength of natural selection increases, whereas the strength of drift decreases, and vice versa for decreasing *s* and decreasing *N*. So far so good for Millstein's account.

However, recall that on Millstein's view, natural selection and drift, when they operate together, are two separate sampling processes and that the effective population size affects not merely the strength of natural selection but also the strength of drift. In contrast, causal relevance affects only the strength of natural selection. Now consider a case where *N* decreases but *s* increases in such a way that the crucial quantity *4Ns* increases. In such a case, because the crucial quantity increases, biologists would predict that natural selection tends to dominate drift. However, on Millstein's account, such a prediction would not be justified: when *N* decreases, the strength of drift increases accordingly. As to the strength of natural selection, it is hard to say whether it increases or not: on the one hand, the strength of natural selection tends to diminish because *N* decreases. On the other hand, it tends to increase because *s* increases. Since the two tendencies oppose each other, one might wonder which one wins out. Even if the fact that the crucial quantity increases suggests that the strength of natural selection ends up increasing, one might still wonder whether it increases greatly enough to outcompete drift. Hence, this is a case where it would be hard to say whether natural selection or drift tends to predominate, even if the crucial quantity increases. To be sure, this is not the only case where a disagreement arises between Millstein and the majority of biologists. A similar argument can be provided in considering a second case where *N* increases but *s* decreases in such a way that the crucial quantity *4Ns*

decreases. In such a case, although most biologists would predict that drift tends to dominate natural selection, Millstein would decline to make such a prediction.

These cases represent a divergence of Millstein's account from a broad consensus among biologists. Such a divergence arises ultimately from Millstein's view that when natural selection and drift operate together, there must be two separate sampling processes. As a result, it is possible for the two separate sampling processes either to increase their strengths at the same time or to decrease their strengths at the same time. When such a possible scenario happens, it would be hard to tell whether natural selection or drift tends to predominate. In contrast, if natural selection and drift, when they operate together, are one single sampling process and if their strengths always counteract each other, then neither is it possible for the two strengths to go up together nor is it possible for them to diminish at the same time. And an increase in the crucial quantity *4Ns* means definitely two things: both an increase in the strength of natural selection and a decrease in the strength of drift, which in turn means that natural selection tends to dominate drift, and vice versa for a decrease in the crucial quantity. It is unclear how, on Millstein's account, two separate sampling processes would be so linked together that there definitely is a trade-off between each other's strengths. Given that the causal relevance affects only the strength of natural selection, such a trade-off would seem impossible.

Now, consider the following statement (*T*):

> (*T*) Whenever natural selection and drift operate together, a change in the strength of natural selection implies an inverse change in the strength of drift, and vice versa.

I would say that any account of natural selection and drift would be dismissed as inadequate if it fails to show that (*T*) is the case. Thus, one problem with Millstein's account is clearly this: it fails to meet the statement (*T*). It turns out once again that the problem with Millstein's account arises ultimately from her view that natural selection and drift are two separate sampling processes.

Admittedly, Millstein's account, *most of the time*, leads to no divergence from biologists' judgment as to whether natural selection or drift tends to predominate. That is because when natural selection and drift operate together among a *real* population of organisms, it seems reasonable to suppose that the degree of causal relevance, which is essential to natural selection, remains *constant*, namely, a fixed nonzero value. Accordingly, both the strength of natural selection and the strength of drift depend exclusively on one single factor, i.e., the effective population size. So although natural selection and drift are two separate sampling processes, there could be a trade-off between their strengths. Still, *theoretically speaking*, Millstein's account is *not* consistent with (*T*).

Now let's return to Brandon's account. Recall that he characterizes drift as any deviation from expected result due to sampling error. Brandon's characterization has an immediate and significant advantage: drift as unrepresentative sampling process is inherently related to the effective population size. The smaller the effective population size is, the more likely it is that drift would be unrepresentative, and accordingly, the larger its strength would be. Although a sampling process is drift if and only if it is unrepresentative, not every representative sampling

process is natural selection. Nonetheless, when natural selection and drift operate together, they are one single sampling process, and if this sampling process is representative, then it surely is natural selection or otherwise drift. It follows that if this sampling process is more likely to be representative, then it is less likely to be unrepresentative, and vice versa. So when natural selection and drift operate together, if there is any change in terms of strength, there surely is a trade-off between each other's strengths. In this sense, Brandon's account meets the requirement (*T*).

As shown above in my argument, the trouble with Millstein's causalist account turns out to be the view that natural selection and drift are, ontologically speaking, two separate sampling processes. A question arises: could Millstein avoid the trouble without undermining her causalist account? In my view, it suffices for Millstein to revise her characterization of drift in such a way that drift ceases to be equivalent to an indiscriminate sampling process. To this end, it suffices that unrepresentative sampling processes count as drift. Note that an indiscriminate sampling process could still count as drift, provided that when it operates together with natural selection, the requirement (*T*) is satisfied. As is evident, if unrepresentative and indiscriminate sampling processes both count as drift, then it is *not* the case that drift and natural selection are necessarily mutually exclusive. And the trouble with Millstein's account would be avoided. Such a revised concept of drift is *de-unified* to the extent that it seems to cover heterogeneous cases. I would say that such a de-unification, instead of showing the revised concept is inappropriate, serves to show what is characteristic of drift. Indeed, drift is an umbrella concept supposed to cover all cases where chance plays a role in a sampling process. Given that chance might play a role in a variety of ways and in a variety of situations, it is no wonder why the concept of drift is un-unified and even un-unifiable. Thus, Brandon's account of drift as unrepresentative sampling process turns out to be helpful to Millstein's account of drift as indiscriminate sampling process, and both could be incorporated into an un-unified account of drift.

3 Could Natural Selection Be a Population-Level Causal Process?

According to Millstein, natural selection is by nature comparative: whenever it acts as a cause of evolution, it impinges on comparative and thus population-level properties (e.g., variation in fitness, frequency of traits) rather than on individual-level properties (e.g., fitness, traits). On this view, natural selection is a population-level causal process. One problem with such a view is that it threatens to make natural selection into a shadow process. Indeed, this problem has been identified by Shapiro and Sober (2007) as follows:

> Natural selection is not a population-level causal process. If it were a sampling process operating at the population-level, it would be, so to speak, a shadow process, an epiphenomenon of individual-level causal processes. (Epiphenomenon problem)

Population-level: Natural selection → change in the frequency of traits

↓(supervene on)

Individual-level: The many causes → individual births and deaths

Fig. 9.2 Individual births and deaths constitute the supervenience base of the change in the frequency of traits

To see how the epiphenomenon problem arises, one should be reminded that each of the population-level properties supervenes on individual-level properties (Shapiro and Sober 2007). Now, at the level of individual organisms, few would deny that there are already many causes for individual births and deaths and that the summation of such individual births and deaths constitutes the supervenience base of the change in the frequency of traits (Fig. 9.2).

If many individual-level causes suffice to make a change in the frequency of traits, then natural selection, which is supposed to be the population-level cause of such a change, threatens to be redundant—a shadow process. Or to put it another way, granting that a change in the frequency of traits is the summation effect of many individual-level causes, if natural selection were the population-level cause of such a change, then the same change would be overdetermined. To avoid overdetermination, one cannot help but annihilate natural selection into a shadow process. Without a satisfactory solution to the epiphenomenon and the overdetermination problems, any claim that natural selection is actually a population-level causal process would seem premature.

Shapiro and Sober (2007) have tried to debunk an argument which, supervenience base being equal, requires natural selection to have extra causal efficacy in order to be a population-level and separate cause (in addition to the many individual-level causes). They hold that such a requirement is too demanding. To meet such a requirement is, in their view, a "mission impossible." I am not sure that Shapiro and Sober are on the right track when they say that no extra causal efficacy is needed to qualify natural selection as a separate cause. It is doubtful that natural selection without extra causal efficacy qualifies as a separate cause. Extra or not, one has to show first of all where the causal efficacy of natural selection lies. And that is the crux of the whole issue: on the one hand, natural selection, operating at the population level, must have causal efficacy; otherwise, it would fade into a shadow process. On the other hand, to avoid the overdetermination problem, natural selection should not have causal efficacy. Here lies a dilemma.

In my view, the way out of the dilemma would be a division of labor between natural selection and individual-level causes. The question is how such a division of labor can be envisaged. Here are some basic ideas: natural selection is part of the cause of a change in the frequency of traits. The many individual-level causes do not suffice for such a change. The many individual-level causes have individual births and deaths as their effects. Although such effects suffice to *sum up* the frequency and *calculate* the way the frequency changes, the many individual-level causes do not suffice to *explain causally* why it changes in the way that it does. Let me explain.

It would be helpful to start with an analogy. Suppose two ethnic groups—call them *N* and *I*, respectively—coexist in a kingdom, ruled unfortunately by a tyrant. An apartheid regime is rigorously maintained in the kingdom. Though the two groups currently have about the same population size, group *I* is known to be much more fertile than group *N*. Fearing that the noble group *N* would be outnumbered by the ignoble group *I*, the tyrant has a purge plan in mind and cannot wait to execute it. He promulgates a law according to which the ratio of population size between group *N* and *I* should remain *x*:1, where *x* represents the number of years after the promulgation of the law. The imperial army is responsible for the enforcement of the law. Since the promulgation of the law, each New Year's Eve, the imperial army kills enough members of group *I* to achieve the right ratio and ensure that the law would not be violated. Each New Year's Eve, the whole population of group *I* is chased by the imperial army into a forest. And then officers of the imperial army compete with each other to hunt as many people of group *I* as possible. The hunting game will not end until the legally required ratio of the year is definitely achieved.

Let's see how such a scenario helps to shed light on the causal role played respectively by the tyrant's law and by the brute action of the imperial army. Consider first the tyrant's law. Apparently, without the tyrant's promulgation of the law, the relative frequency in population size between group *N* and *I* would not have changed in accordance with the ratio *x*:1. In this sense, it seems safe to say that each annual change in the relative frequency is caused by the tyrant's law.

Now consider a member of group *I*, say, Ian, who unfortunately was hunted down and fell victim to an officer of the imperial army, say, Ned. What is the cause of Ian's death? Undoubtedly, Ian was killed by Ned. So without Ned's killing, Ian would not have died. In this sense, Ned's killing causes Ian's death. But didn't Ian get killed also by the tyrant's law? That is, doesn't the tyrant's promulgation of the law also cause Ian's death? The answer is apparently yes, for without the law, Ned would not have killed Ian. Few, if any, would deny that the tyrant's law is part of the cause of Ian's death. Indeed, to tell a causal story about Ian's death, one needs to invoke not only Ned's brutal behavior but also the tyrant's enactment of the law. Otherwise, the causal story would be regretfully incomplete.

However, the fact that the two factors contribute jointly to Ian's death shouldn't mislead us into conflating their separate effects. *A definite number of people* among group *I* are killed due to the tyrant's law, while it is Ned's killing that makes *Ian* one of them. The tyrant's law determines *how many* people will die, while each officer of the imperial army determines exactly *who* are going to be victims. What is the separate effect of the tyrant's brutal law? Answer: the (change in the) relative frequency in population size between group *N* and *I*. And what is the separate effect of Ned's killing? Answer: Ian's being among the dead. From the combination of the two separate effects, one gets the result: Ian's death. So the fact that both the tyrant's law and Ned's killing are part of the causal story of Ian's death does not imply that either of them fails to have a separate effect. Indeed, the tyrant's law is causally responsible for a definite number of people's death, while Ned's killing is for Ian's being among one of them. To causally explain Ian's death, one needs to pinpoint not merely Ned's brutal killing but also the tyrant's terrifying law.

If the preceding analysis is correct, then we have a ready answer to the epiphenomenon problem. Suppose someone raises the epiphenomenon problem in the following way:

One may wonder: to explain why the relative frequency in population size between group *N* and *I* changes in accordance with a yearly increasing ratio, is it really necessary to invoke the tyrant's law? It seems that annual genocides, committed by the imperial army, suffice to explain why the relative frequency changes in a yearly increasing way. If one has each and every causal story about how, each year, each officer of the imperial army kills his victims and how survivors escape death, then such many causal stories, in addition to a little statistics, suffice to show why the relative frequency changes in a yearly increasing way. So the tyrant's law is causally redundant: it has no effect at all. The change in the relative frequency is not caused by the law. Rather, it is a summation result due to many causes that one can only find in each gory and detailed causal story.

The epiphenomenon problem, in this case, fails to cast doubt on the causal efficacy of the tyrant's law. Few, if any, would deny that the tyrant's law is part of the cause of each annual change in the relative population size between group *N* and *I*.

On the other hand, though the many detailed causal stories, in addition to a little statistics, suffice to *sum up* the annual relative frequency and *calculate* the way the relative frequency changes, they do not suffice to *explain causally* why it changes in the way that it does. A complete explanation of why the relative frequency changes in a yearly increasing way must invoke the tyrant's law. The distinction between "suffice to sum up" and "suffice to explain causally" permits an answer to the overdetermination problem: the annual change in the relative frequency is not overdetermined, for the many detailed causal stories, along with statistics, do not suffice to explain causally why the relative frequency changes in a yearly increasing way.

As mentioned above, the view that natural selection is a population-level cause of the change in the relative frequency among a population from one generation to the next faces a dilemma: it gets into trouble either with the epiphenomenon problem or with the overdetermination problem. However, with respect to causal efficacy, if the tyrant's law can serve as an analogy for natural selection, then a way out of the dilemma is accessible to natural selection. So my argument, if successful, would be good news for the population-level cause view of natural selection. More specifically, it would be an argument in support of Millstein's view that natural selection is a population-level causal process in the sense that it acts as a cause of evolution and that it impinges on comparative and thus population-level properties.[4] The fact that natural selection operates at the population level provides no support for it being a shadow process. Despite appearances, natural selection is part of the cause of the change in the relative frequency.

[4] I leave open an interesting question as to whether natural selection, operating at the population level, is as causally robust as the many events occurring at the individual level. Stuart Glennan (2009) distinguishes causal relevance and causal productivity and argues that at the population level of natural selection, there can be causal relevance without causal production. In response, Millstein (2013) argues that there *is* causal production at the population level of the natural selection process.

4 Conclusion

Selection-of and selection-for, arguably, are one and the same sampling process. Thus, an adequate account of natural selection should be broad enough to include both selection-for and selection-of. Millstein's causalist account, however, defines discriminate sampling process (natural selection) in terms of selection-for only. Her view would be more complete if the notion of discriminate sampling process includes both *F*-discriminate (selection-of) and *C*-discriminate (selection-for) sampling processes.

A second problem with Millstein's causalist account is that it does not meet a plausible requirement: namely, whenever natural selection and drift operate together, a change in the strength of natural selection implies an inverse change in the strength of drift, and vice versa. In contrast, Brandon's account makes perfect sense of the trade-off between the two factors' strengths. The problem with Millstein's account arises from the view that, ontologically speaking, natural selection and drift are two separate sampling processes. To avoid the problem, it suffices for Millstein to de-unify the concept of drift in such a way that drift comprises both unrepresentative and indiscriminate sampling processes. Since drift is an umbrella concept supposed to cover all cases where chance plays a role in a sampling process, such an un-unifying (even un-unifiable) drift concept, of which Brandon's and Millstein's account of drift are a part, turns out to be credible rather than ad hoc.

The view that natural selection is a population-level cause seems to face a dilemma: it is undermined either by the epiphenomenon problem or by the overdetermination problem. The prospects of the view thus turn on a satisfactory solution to both the problems. With the help of an analogy, I show how the two problems can be dealt with: natural selection is not a shadow process, for it is part of the cause of a change in the relative frequency of traits. On the other hand, the overdetermination problem can be avoided, for although many individual-level causes suffice to sum up and calculate how the relative frequency changes, they do not suffice to causally explain why it changes in the way that it does.

Millstein's causalist account of natural selection and drift is instructive and persuasive as well. Still, one may wonder whether it is too incomplete to be adequate. It is noteworthy that the idea of fitness plays no role in Millstein's characterization of natural selection and drift. To be sure, Millstein has good reason to refrain from using the term "fitness": it is contentious, and she does not need it to make the points she wants to make. Nonetheless, it seems undeniable that biologists routinely use (variation in) fitness to account for evolutionary events. Though it remains unclear whether an explanation in terms of (variation in) fitness would properly be called causal, it seems safe to say that any account of natural selection and drift, in which fitness plays no significant role, would hardly be adequate. Indeed, any such account would seem to throw out the baby with the bath water. How to put fitness into the causalist account and give it due weight? An answer to this question must await, however, another paper.

Acknowledgements I am extremely grateful to Roberta L. Millstein for comments that led to substantial improvements in the final version of this chapter. I would like to thank an anonymous referee who helped me to clarify the issues discussed here. I am grateful to Hsiang-Ke Chao and Szu-Ting Chen for organizing the wonderful conference. I gratefully acknowledge support from the National Science Council, Taiwan (NSC98-2410-H-031-002-MY3).

References

Beatty, J. 1984. Chance and natural selection. *Philosophy of Science* 51: 183–211.

Bouchard, F., and A. Rosenberg. 2004. Fitness, probability, and the principles of natural selection. *The British Journal for the Philosophy of Science* 55: 693–712.

Brandon, R. 2005. The difference between selection and drift: A reply to Millstein. *Biology and Philosophy* 20: 153–170.

Brandon, R. 2006. The principle of drift: Biology's first law. *Journal of Philosophy* 102: 319–335.

Brandon, R. 2010. Natural selection. In *The Stanford encyclopedia of philosophy (Fall 2010 Edition)*, ed. Edward N. Zalta. URL: http://plato.stanford.edu/archives/fall2010/entries/natural-selection/

Brandon, R., and S. Carson. 1996. The indeterministic character of evolutionary theory: No "no hidden variables proof" but no room for determinism either. *Philosophy of Science* 63: 315–337.

Glennan, S. 2009. Productivity, relevance, and natural selection. *Biology and Philosophy* 24: 325–339.

Hartl, D.L., and A.G. Clark. 1989. *Principles of population genetics*. Sunderland: Sinauer Associates.

Matthen, M., and A. Ariew. 2002. Two ways of thinking about fitness and natural selection. *Journal of Philosophy* 99: 55–83.

Millstein, R.L. 2002. Are random drift and natural selection conceptually distinct? *Biology and Philosophy* 17: 33–53.

Millstein, R.L. 2005. Selection vs. Drift: A response to Brandon's reply. *Biology and Philosophy* 20: 171–175.

Millstein, R.L. 2006. Natural selection as a population-level causal process. *The British Journal for the Philosophy of Science* 57: 627–653.

Millstein, R.L., and R.A. Skipper Jr. 2007. Population genetics. In *The Cambridge companion to the philosophy of biology*, ed. D. Hull and M. Ruse, 22–43. Cambridge: Cambridge University Press.

Millstein, R.L. 2013. Natural selection and causal productivity: A reply to Glennan. In *Mechanism and causality in biology and economics*, ed. Hsiang-Ke Chao, Szu-Ting Chen, and Roberta L. Millstein, 147–163. Dordrecht: Springer.

Shapiro, L., and E. Sober. 2007. Epiphenomenalism—The do's and the don'ts. In *Studies in causality: Historical and contemporary*, ed. G. Wolters and P. Machamer, 235–264. Pittsburgh: University of Pittsburgh Press.

Sober, E. 1984. *The nature of selection*. Cambridge, MA: MIT Press.

Walsh, D.M., T. Lewens, and A. Ariew. 2002. The trials of life: Natural selection and random drift. *Philosophy of Science* 69: 452–473.

Part IV
Across Boundaries Between Biology and Economics

Chapter 10
Mechanisms and Extrapolation in the Abortion-Crime Controversy

Daniel Steel

Abstract John Donohue and Steven Levitt's seminal and controversial article, *The Impact of Legalized Abortion on Crime*, famously argues that the legalization of abortion in 1973 in the Unites States is a significant factor explaining the surprising decline in crime rates that occurred there in the 1990s. In this chapter, I examine the role of extrapolation in Donohue and Levitt's study and draw three main philosophical conclusions. First, several different types of causal claims might be at issue in an extrapolation—including claims about mechanisms and probabilistic causal effects—and these distinctions matter for methodology because different conditions may be required to support extrapolation in each case. Secondly, scientific study of a phenomenon typically generates evidence at a variety of levels of aggregation, and this has important implications for extrapolation. The third and final point follows on the heels of the second. Like almost all other scientific inferences, extrapolations are normally components of a complex web of interrelated evidence that must be considered together in assessing a hypothesis.

John Donohue and Steven Levitt's (2001) seminal and controversial article, *The Impact of Legalized Abortion on Crime*, famously argues that the legalization of abortion in 1973 in the Unites States is a significant factor explaining the surprising decline in crime rates that occurred there in the 1990s. Their argument is interesting from a philosophical perspective for a number of reasons, one of which is its use of a variety of methodological approaches in building a case for a causal claim. Although most of the debate surrounding Donohue and Levitt's hypothesis has focused on the interpretation of the statistical data, an important part of their case involves tracing a mechanism whereby legalization of abortion could reduce crime rates. The central idea is that unwanted children—whose births were likely to have

D. Steel (✉)
Department of Philosophy, Michigan State University, 503 South Kedzie Hall, 368 Farm Lane, East Lansing, MI 48824-1032, USA
e-mail: steel@msu.edu

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_10,

been prevented by abortion were it available—are more likely to be born into circumstances that increase the chance that they will engage in criminal behavior upon entering early adulthood. One line of evidence cited by Donohue and Levitt for this hypothesis consists of data from Scandinavia and Eastern Europe, where for some periods of the twentieth century, women desiring an abortion were required to receive legal permission. Thus, data concerning the life outcomes of children whose mothers requested but were denied abortions are directly relevant to the mechanism proposed by Donohue and Levitt. The use of data from Scandinavia and Eastern Europe to support a claim about a mechanism in the United States is an example of an *extrapolation*, that is, using a causal relationship found in one context as a basis for an inference about causal relationships in another that may differ in a number of relevant respects. Extrapolation is essential for imparting significance to scientific research beyond the confines of the original investigation, yet relatively little philosophical attention has been devoted to the question of how and under what circumstances such inferences are justifiable.

In this chapter, I examine the role of extrapolation in Donohue and Levitt's study from the perspective of some recent attempts to clarify the underlying logic and principles of such inferences (Pearl and Bareinboim 2011; Steel 2008). This project is intended both to explicate the methodology of the Donohue and Levitt study as well as to advance philosophical understanding of extrapolation. Three main philosophical themes emerge from the discussion below. First, several different types of causal claims might be at issue in an extrapolation—including claims about mechanisms and probabilistic causal effects—and these distinctions matter for methodology because different conditions may be required to support extrapolation in each case. Secondly, scientific study of a phenomenon typically generates evidence at a variety of levels of aggregation, and this has important implications for extrapolation. The Donohue and Levitt study, for example, discusses data concerning psychosocial effects of unwantedness on individual children in addition to comparisons between rates of abortion and crime among different states in the USA. As explained in the final section, data from the macro-level can provide a means of indirectly testing assumptions about similarities of the model and target made at the level of mechanisms. The third and final point follows on the heels of the second. Like almost all other scientific inferences, extrapolations are normally components of a complex web of interrelated evidence that must be considered together in assessing a hypothesis. Thus, to focus on whether extrapolation *alone* could have established a conclusion in a real scientific example (see Lafollette and Shanks 1996) would be, more often than not, to miss the point. In real-life cases, the question can only be whether and to what extent the extrapolation strengthens the overall body of evidence.

I begin with a synopsis of the Donohue and Levitt study. Next in Sect. 2, I consider the role of multiple levels in the study, in particular, the connection between the psychosocial mechanism concerning unwantedness and criminal behavior and the state-level comparisons concerning abortion and crime rates. I introduce the notion of a "scale-up model" that links a micro-mechanism to a macro-level statistical relationship. In Sect. 3, I present a framework for

conceptualizing extrapolation that combines elements from Steel (2008) and Pearl and Barienboim (2011) and explain its relevance to the Donohue and Levitt study. Section 4 explains how macro-level data can be used to support an extrapolation made at the level of mechanisms and links this to the final of the three philosophical themes listed above.

1 Donohue and Levitt's Argument

In this section, I examine the structure of Donohue and Levitt's argument for their hypothesis that the legalization of abortion in 1973 is the primary cause of the abrupt, nationwide, and persistent drop in crime rates that began in the early 1990s in the United States. Their case can be usefully organized into three main intertwined sub-arguments. First, they argue that the decline in crime rates is not adequately explained by alternative causes, such as increased incarceration or the decline of the crack epidemic. Secondly, tracing a mechanism through which legalized abortion in 1973 could result in a drop in the crime rate 18–24 years later when the first post-legalization birth cohort entered its peak crime age. Finally, a statistical argument focused on differing timing of declines in crime rates between earlier and later legalizing states and on the timing of which types of crime dropped first.

Nationwide statistics of rates of violent and property crimes in the USA peaked in the first half of the 1990s and then began a steady decline that has continued into the first decade of the new century (Truman and Rand 2010). The abrupt, widespread, and persistent nature of this decline constrains possible explanations. Any explanation must be able to explain why the decline began when it did, why it occurred throughout the USA, and why it has persisted, now, for over 15 years. Donohue and Levitt's hypothesis directly explains each of these features. First, the impact of abortion legalization on crime would begin to be felt approximately 18 years after the Roe v. Wade decision in 1973—that is, in 1991—when the first post-legalization cohort entered its peak crime years. Secondly, Roe v. Wade invalidated legal bans on abortion in all jurisdictions in the USA and hence is a cause of nationwide scope. Finally, abortion legalization is a cause that exerts its impact over an extended period of time. The effect of legalization would continue to be felt as prelegalization generations age out and post-legalization generations age in. Thus, Donohue and Levitt predict that the crime-reducing impact of legalized abortion would continue to be felt until around 2020 (2001, p. 415).

Donohue and Levitt argue that a number of alternative explanations of the decline fail to account for one or more of the three basic features listed above. For example, consider one of the most popularly cited explanations, innovative policing strategies. These were instituted in New York City only after the decline in crime had already begun there and were not implemented in many other cities, such as Los Angeles, that also experienced significant crime reductions (Levitt 2004, pp. 172–173). Another explanation points to the crack cocaine epidemic, which struck

many American cities in the late 1980s and subsided in the 1990s. However, the subsidence of the crack epidemic does not explain drops in crime in suburban and rural areas in which crack was never a serious problem and fails to explain why the crime would persistently fall below pre-crack levels (Donohue and Levitt 2004; Fryer et al. 2005; Levitt 2005). In addition, although increased incarceration and swelling ranks of police forces likely played a role in reducing the nationwide crime rate, these factors predate the decline in crime in the 1990s by at least a decade (Donohue and Levitt 2001, p. 380; Levitt 2004). Since Donohue and Levitt's (2001) essay, a few new explanations of the surprising 1990s crime decline have been proposed. One explanation that accounts for all three basic features of the decline focuses on the phase out of leaded gasoline required by the 1970 Clean Air Act (Reyes 2007). Early childhood lead exposure is known to cause a number of cognitive deficits, some of which are related to aggressive and violent behavior in adulthood. Thus, environmental regulations enacted by the federal government that sharply reduced lead exposure could have had a crime-diminishing effect a generation later, in much the same manner as legalized abortion according to Donohue and Levitt's hypothesis. The lead-crime hypothesis also has the advantage of explaining rising crime rates in the two decades prior to the 1990s, which could have partially resulted from increased lead exposure throughout the mid-twentieth century to the early 1970s (Reyes 2007, p. 33). Of course, the causes surveyed above are not mutually exclusive, and the most likely scenario is that fluctuations in crime rates are due to a variety of factors, including some not mentioned here.[1]

The next part of Donohue and Levitt's case consists of tracing the mechanism from legalized abortion to reduced crime (section III of their 2001 article). They in fact propose two mechanisms through which legalized abortion could reduce the crime rate: cohort-size reduction and selection. Cohort-size reduction is the idea that legalized abortion would reduce the birth rate, which in turn would mean a smaller cohort of individuals aging into the high-crime years of 18–24 in the early 1990s. Selection is a more interesting and controversial mechanism and is the one that will occupy our attention here.

In general, the term "selection effect" refers to a situation in which a group of individuals defined in an inquiry is not a random sample of the population but instead differs in some further way that is relevant to topic under investigation. In this context, the selection mechanism operates if children born to mothers who wished to terminate their pregnancy are not a random sample of births generally but are much more likely to be born into adverse family, social, or economic circumstances that put them at greater risk for criminal activity later in life. The selection mechanism can be helpfully further divided into two separate sub-mechanisms. The first concerns the impacts of being born unwanted on the child's

[1] Reyes (2007, p. 36) regards both reduced lead exposure and legalized abortion as significant factors. Levitt (2004) critically examines several further proposed explanations of the crime drop, including demographic factors and improved economic conditions. Wadsworth (2010) proposes that immigration may have contributed to the 1990s decline in crime rates.

psychological and social development, while the second has to do with factors that would make the pregnancy unwanted in the first place. To illustrate the first of these two sub-mechanisms, a mother might resent a child resulting from an unwanted pregnancy and thus have a more negative and less affectionate attitude towards it, which could exert a damaging psychological effect on the child. To illustrate the second, the mother might desire to terminate the pregnancy because she is an unmarried teenager, and children born to unwed teens, regardless of whether they are wanted or unwanted, may be at greater risk for a number of adverse life outcomes, including criminality.

The most direct empirical test of the selection mechanism would identify children whose mothers desired to terminate their pregnancies but were prevented from doing so by legal restrictions on abortion. Data sets of this kind exist for several European countries wherein, during the 1930s–1960s, women desiring to terminate a pregnancy were required to file an application to obtain legal permission to do so. Thus, cases of women whose applications for abortion were denied and who subsequently gave birth constitute precisely the type of sample in question. Samples of this kind exist for Sweden and the former Czechoslovakia and have been the basis for a number of studies documenting the psychological and social effects on children of being born unwanted (see David et al. 1988). The most thorough of these studies concerns the Prague cohort of 220 children born to women whose request for abortion was twice denied, on the initial application and then on appeal. This cohort is known as UP (for unwanted pregnancy) and was matched with 220 AP (accepted pregnancy) control children. Since the Prague study was designed to test of the effects on the child of being unwanted (i.e., the first sub-mechanism mentioned above), the UP and AP subjects were matched on socioeconomic terms as well as on a number of other factors such as family size. In addition to collecting medical, school, and legal data, the study conducted double-blind interviews of parents, teachers, and children, which for the latter group included psychological and intelligence tests. The study focused on several age points: birth, age 9, ages 14–16, and ages 21–23. The study found no physiological or health differences between the UP and AP children at birth but a consistent pattern of less favorable outcomes in the subsequent follow-ups. For example, at age 9, UP children were significantly more likely to be rejected by peers (Matejcek et al. 1988, pp. 69–70) and to have difficulty in adapting adequately to frustration (ibid. 70–71). By the age 14–16 follow-up, the gap in school achievement between UP and AP children had become statistically significant (ibid. 88). By the age 21–23 follow-up, more than twice as many UP subjects had been sentenced in court, and the average prison term of those sentences for the UP subjects was more than double the average sentence of the AP controls (Dytrych et al. 1988, p. 94). Since UP and AP families were matched for socioeconomic status, most of these differences appear to be due to the less favorable internal family dynamics of the UP subjects (Matejcek et al. 1988, pp. 74–75).

The European studies are cited as support for the selection mechanism in both the original Donohue and Levitt article (2001, p. 388) as well as in a summary of that argument given by Levitt (2004, p. 182). The importance of these studies in

arguing for the selection mechanism is easy to appreciate given that there are no known records in the USA of women who desired to terminate pregnancy but were denied legal right to do so.[2] However, there are studies using USA data that bear on the selection mechanism less directly. For example, Donohue and Levitt cite studies finding that women who terminate pregnancies are, not surprisingly, significantly more likely to be in circumstances known to adversely affect the life prospects of children, such as being an unmarried teenager (Levine et al. 1999). Similarly, it is possible to study overall effects of abortion legalization on the well-being of birth cohorts. Gruber et al. (1999) do this and find that post-1973 birth cohorts exhibited a marked reduction in a number of adverse factors, including single-parent households and poverty.

In addition to tracing a mechanism from legalized abortion to the 1990s decline in crime rates, Donohue and Levitt argued for their hypothesis on the basis of statistical data concerning abortion and crime rates. Donohue and Levitt's statistical argument turns on three main points. First, since five states in the USA (Alaska, California, Hawaii, New York, and Washington) legalized abortion around 1970, Donohue and Levitt's hypothesis predicts that the declines in the early legalizing states began about 3 years before those in the other states. And since abortion rates continued to vary between states after Roe v. Wade, Donohue and Levitt's hypothesis predicts that states with higher abortion rates would experience greater reductions in crime. Secondly, because teenagers are much more likely to commit property crimes than violent crimes, Donohue and Levitt's hypothesis predicts that the decline in crime rates would begin with property crime and then spread to more violent crimes. Finally, given the "age in" impact of abortion legalization, Donohue and Levitt's hypothesis predicts that the decline in crime would disproportionately result from cohorts born after legalization. Donohue and Levitt argue that the data support all three of these predictions.[3]

Besides its intimate link to a perennial "hot button" social issue, the above complex argument makes a fascinating case study in social science methodology. Numerous philosophers and social scientists have discussed the role of mechanisms in providing support for causal claims in social science (Elster 1989; George and Bennett 2005; Hedstrom and Swedberg 1999; Kincaid 1996; Little 1992, 1998; Reiss 2007; Steel 2004). Two features of the Donohue and Levitt study make it interesting and relevant in connection to these discussions: (1) the interconnected role of mechanisms and macro-level statistical data in supporting its central hypothesis and (2) the role of extrapolation in providing evidence for a mechanism.

[2] No doubt such women exist. But since there was never any official process of applying for permission to terminate a pregnancy in the USA, there is no way to identify which women these were.

[3] Several critics have challenged these statistical arguments (Joyce 2003; Foote and Goetz 2008). See Donohue and Levitt (2004, 2008) for replies.

2 Mechanisms and Scale-Up Models

Tracing a mechanism can strengthen a causal inference in several ways. It allows for additional tests of the hypothesis and creates further lines of relevant evidence. In addition, experimental or quasi-experimental studies of important aspects of the mechanism may be possible even when no such studies are feasible at the macro-level (Steel 2011). These points are illustrated in the Donohue and Levitt, wherein studies of women who requested but were denied abortion constitute "natural" experiments of the effects of being born unwanted on child development. But in order for inquiries concerning mechanisms to support macro-level causal claims, there must be some reasonably clear connection between the mechanism, which typically involves individual behaviors and interactions, and the macro-level, which makes general claims about a population. What I will call a "scale-up model" is used to forge a link between micro and macro. A scale-up model specifies how an inference can be made from mechanisms to macro-level phenomenon. Donohue and Levitt in fact explicitly present a scale-up model, which they describe as a "back-of-the-envelope" calculation that provides a "crude prediction of the impact of legalized abortion on crime" (2001, p. 389).

Their approach combines research on how legalized abortion affects the composition of birth cohorts as judged by four factors—race, teenage motherhood, unmarried motherhood, and unwantedness—along with research on the impact of each of those factors on criminality (ibid). The model then breaks down 1990 Census data using the eight possible combinations of the first three factors (apparently assuming that race is either white or black), finding the proportion of the population in each group. They then use estimates from a study of the impact of Roe v. Wade on birth rates (Levine et al. 1999) to decide what those proportions would have been if abortion had not been legalized. Next, Donohue and Levitt use previous research to assign crime rates for each cell (e.g., for children of a white unmarried teenage mother). Thus, the effects of abortion on crime mediated by the first three factors can be estimated by summing the proportion-weighted crime rates for each cell with two different proportion weightings: one based on 1990 Census data and the other set based on estimates of what these proportions would have been if abortion had not been legalized. Since unwantedness is not measured in the data, Donohue and Levitt estimate the number of unwanted births by assuming that 75 % of unwanted births would be prevented by abortion.[4] They then extrapolate the result that children born from unwanted pregnancies are about twice as likely to

[4] That is, number of abortions ÷ 75 % = number of unwanted births (if abortion had not been legalized). Thus, given the number of abortions (for which there is data), the number of unwanted births in the hypothetical nonlegalization scenario can be estimated. The number of unwanted births in the actual legalization case would just be this number minus the number of additional abortions performed due to Roe v. Wade.

commit crimes.[5] For homicide, this rough model predicts an 11 % reduction due to legalized abortion by 1997, when homicide rates had fallen from their peak by 30–40 % (Donohue and Levitt 2001, p. 390).

This scale-up model is relevant to the present discussion in several respects. It explicitly involves an across-population extrapolation, as the estimate for the impact of unwantedness on criminal behavior is directly borrowed from the European studies described above. In addition, the scale-up model is important for understanding the relation of extrapolation to multiple levels of aggregation in making an overall case for a causal claim. In particular, the scale-up model provides an indirect way to test the claim that the populations are sufficiently similar to justify the extrapolation. This point is discussed in greater detail below in Sect. 4.

3 A Conceptual Framework for Extrapolation

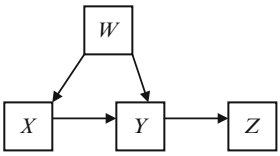
Transferring or extrapolating results across populations is essential for making scientific inquiry relevant to practical problems. Reasoned answers to questions about the causes of the decline in crime rates in the 1990s, or about the health impacts of a pesticide, and many other issues must consider a variety of studies whose data sets are not all drawn from the same population. The role of the European studies on the impacts of unwantedness is the most obvious but not sole example of this in the Donohue and Levitt study. In tracing the mechanism they propose, Donohue and Levitt cite studies—for example, about the effects of teenage motherhood on the life outcomes of the child—that use distinct US data sets. In this section, I synthesize and further develop approaches to extrapolation proposed by myself (Steel 2008) and Pearl and Bareinboim (2011).

3.1 Selection Diagrams and a Definition of Extrapolation

I will use the term *model* to refer to the population that is the basis of the extrapolation and the term *target* for the population that is the object of the extrapolation. In the Donohue and Levitt study, the European women in Sweden and Czechoslovakia during times in which terminations of pregnancies required legal approval are the models and the USA is the target. As this example illustrates, there may be more than one model. Extrapolating (or transporting, to use the term favored by Pearl and Bareinboim) depends on some background knowledge about ways in which the model and target are likely to differ and ways in which they are

[5] Note that this last assumption is consistent with the results of the Prague cohort study described above. However, Donohue and Levitt cite a study concerning a Finnish data set here (Rasanen et al. 1999).

Fig. 10.1 A DAG



likely to be similar (Steel 2008, pp. 88–89). Typically, the model does not resemble the target in all relevant respects, and hence extrapolation can involve some adjustment for to account for differences. A first step towards a conceptual framework for representing extrapolation, then, is a means for compactly and perspicuously representing causally relevant similarities and differences between the model and target. Steel (2008) and Pearl and Bareinboim (2011) develop similar representational frameworks for this purpose through an extension of directed acyclic graphs (DAGs), which are often used to represent causal relationships. A DAG consists of a set of nodes linked by arrows, for example, as in Fig. 10.1. A DAG is *directed* in that every line (or "edge") has an arrowhead attached at one end, and it is *acyclic* in that it does not contain loops, that is, a sequence of arrows aligned head to tail that begin and end with the same node. If there is an arrow pointing directly from node X to node Y, then X is said to be a *parent* of Y. For example, X is a parent of Y in Fig. 10.1, and Y is a parent of Z. If Y can be reached from X by following a chain of arrows aligned head to tail, then Y is said to be a *descendant* of X.[6] Thus, in Fig. 10.1, X, Y, and Z are all descendants of W. The nodes in a DAG are normally taken to represent variables. When DAGs are interpreted causally, parents are direct causes and descendants are effects. A Bayesian network consists of a DAG together with a probability distribution that satisfies something known as the Markov condition with respect to that DAG. The Markov condition asserts that every variable in the DAG is probabilistically independent of its non-descendants conditional on its parents. For example, in Fig. 10.1, Z is probabilistically independent of W and X conditional on Y. Intuitively, this means that once the value of the variable Y is known, learning the values of W and X provides no further information concerning the value of Z. The graphical concept of d-separation allows one to read off all of the independence relationships entailed by the Markov condition for a DAG (see the Appendix 1 for the "Definition of d-Separation"). The Markov condition will play an important role in the account of extrapolation described below.

A simple extension of DAGs can be used to represent similarities and differences between causal relationships in model and target populations. That extension consists of adding additional variables to a DAG to represent differences between model and target populations that may alter the relationships represented in the DAG. For example, consider the diagram in Fig. 10.2. This diagram represents a

[6] This definition should be understood to entail that every node is descendant of itself (as any node is trivially reachable from itself). This seemingly odd feature of the definition simplifies the statement of the Markov condition.

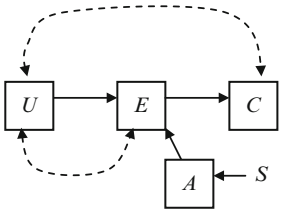


Fig. 10.2 A selection diagram with an unmeasured common cause

hypothesis about the causal relationship between unwantedness and crime. The variable U indicates whether or not the person was born from an unwanted pregnancy, E is a variable indicating harmful psychological effects (e.g., an impaired ability to adapt to frustration), A indicates whether or not the child was adopted, and C indicates whether the person has been convicted of a crime. In contrast, S is a variable that represent unmeasured factors that may create differences between the two populations. In Steel (2008, pp. 58–62) these are called *disrupting factors*, while in Pearl and Bareinboim (2011, p. 6) they are called *selection variables*. I follow Pearl and Bareinboim's terminology here, as the term "disruption" suggests factors that entirely block a causal relationship, while the differences between model and target could come in other forms. To understand selection variables, it is important to realize that two causes may interact with one another in bringing about an effect. For example, by altering the adoption rate, S may change the impact of U upon E. That is, a child born from an unwanted pregnancy but adopted into a loving family shortly after birth would presumably be spared the deleterious psychological effects of unwantedness. In the extreme case, if every child born unwanted were adopted, the harmful effects of unwantedness might be eliminated entirely. That extreme scenario seems rather improbable—a survey of studies concerning children born from unwanted pregnancies found a maximum adoption rate of around 20 % (Dagg 1991, p. 582)—but the important point is differences in the selection variable S could mitigate the deleterious psychological effects of unwantedness. Note that the absence of selection variables pointing into variables other than A in Fig. 10.2 is also significant. For instance, the diagram in Fig. 10.2 says that E impacts C in the target exactly as in the model.

Pearl and Bareinboim (2011) refer to graphs like the one in Fig. 10.2 as selection diagrams. Selection diagrams, then, represent judgments about similarities and differences between model and target populations. A selection variable indicates a source of potential difference between the model and target. For example, the selection diagram alerts us to the possibility that the effect of unwantedness upon psychological difficulties in the target may differ from that in the model due to differences in adoption rates between the two populations. A selection diagram, then, represents the causal structure in the target, namely, the DAG that results from removing the selection variables. In addition, the selection diagram indicates ways in which the causal structure in the target may differ from the model. For example, in Fig. 10.2, it is possible that the selection variable changes the distribution of A in such a way as to eliminate all influence of U upon E in the target. Even if the causal structure represented by the DAG is the same for model and target, the quantitative

causal relationships may differ as a result of the selection variables. A key feature of a selection diagram is that the selection variables explain all of the differences between the probability distributions in the model and target. Letting P and P^* be the probability distributions in the model and target, respectively, this means that any P^* probability is equal to the corresponding P probability conditional on the set of selection variables, S. This is particularly important when a probabilistic causal relationship can be estimated in the model but not in the target (e.g., because an experimental intervention was performed in the former but not in the latter). *Causal effects* are one important type of probabilistic causal claim. The causal effect of X upon Y is the probability distribution Y conditional on an intervention on X (see Pearl 2000, p. 70). Pearl uses the "do-operator," written as "$do(x)$," to indicate that the value of the variable X has been set by an intervention rather than passively observed, so that the causal effect of X upon Y would be written as $P(y \mid do(x))$.[7] Thus, if Fig. 10.2 is the correct selection diagram, $P^*(c|do(u)) = P(c|do(u), s)$, that is, the causal effect of U on C in the target is equal to the corresponding causal effect in the model conditional on the selection variable S. Since selection variables are assumed to be unmeasured, it may not be possible to estimate $P(c|do(u), s)$ directly from data drawn from the model population. Consequently, extrapolating a causal effect from the model to target requires some means of reducing $P(c|do(u), s)$ to a formula in which $do(u)$ and s never occur in the same probability. In the subsequent section, we will consider an example of how this can work. Pearl and Bareinboim's selection diagrams also include dashed-double-headed arrows to represent the presence of unmeasured common causes (as in Fig. 10.2).[8] As a result of the unmeasured common cause of U and E, the causal effect of U upon C cannot be identified from observational data in the target (see Pearl 2000, p. 94).

I now use selection diagrams to define extrapolation and integration. The definition of extrapolation and direct extrapolation mostly parallel those given in Pearl and Bareinboim (2011, p. 9) but diverge from them in one important respect that I explain below. The definition of integration is original and useful for the thinking about the Donohue and Levitt study.

Definition 10.1 (Extrapolation). Let Π be the model population and Π^* the target characterized by the probability distributions P and P^*, respectively, and let D be a selection diagram relating Π and Π^*. Then a causal relation R *can be extrapolated from* Π *to* Π^* if and only if $R(\Pi^*)$ is identifiable given the conjunction $R(\Pi)$, P, P^*, and D.

Extrapolation is *direct* when $R(\Pi)$ is the same as $R(\Pi^*)$. When extrapolation is direct, no modification or adjustment to the causal relationship estimated in the model is needed; it transfers as is to the target. As will be illustrated in the

[7] Here I follow the convention of having lower-case letter represent particular values of the variables represented by the corresponding upper-case letters.

[8] DAGs with double-headed arrows representing unmeasured common causes are known as semi-Markovian models.

subsequent section, extrapolation need not be direct and can involve adjustments to the causal relationship found in the model.

A few clarifications of Definition 10.1 are in order. First, for simplicity, Definition 10.1 is limited to the case in which there is only one model. The generalized version of the definition allowing for multiple model populations Π_1 through Π_n would have corresponding probability distributions P_1 through P_n, as well as selection diagrams D_1 through D_n (as each model could differ from the target its own way). A further generalization of the definition would allow for distinct causal relationships to be extrapolated from each model. The result of this generalization is more properly regarded as a definition of *integration* rather than extrapolation, since it involves the combination of a number of studies performed on several populations.

Definition 10.2 (Integration). Let Π_1 through Π_n be model populations characterized, respectively, by the probability distributions P_1 through P_n, and let Π^* be the target population, with probability distribution P^*. Let D_1 through D_n be selection diagrams for the pairs $\langle \Pi_1, \Pi^* \rangle$ through $\langle \Pi_n, \Pi^* \rangle$, respectively. Then causal relations $R_1(\Pi_1)$ through $R_n(\Pi_n)$ *can be integrated to learn* $R(\Pi^*)$ if and only if $R(\Pi^*)$ is identifiable given conjunction of $R_1(\Pi_1)$ through $R_n(\Pi_n)$, P_1 through P_n, P^*, and D_1 through D_n.

This definition pertains to cases in which results from a number of studies of disparate model populations are combined to infer a potentially new causal relationship in the target. For instance, Donohue and Levitt's scale-up model described in Sect. 2 integrates a number of distinct results from studies performed in several populations in order to form a rough estimate of a causal relationship not studied in any of them, namely, the effect of legalized abortion in 1973 on crime in the 1990s in the USA. Notice that Definition 10.1 is a special case of Definition 10.2 in which there is only one model and the causal relationship to be inferred in the target is the same one as that in the model.

A second clarification concerns the causal relation R. There are in fact several types of causal relationships one might wish to extrapolate. Pearl and Bareinboim (2011) focus on extrapolating causal effects. Steel (2008) is primarily concerned with extrapolation of positive causal relevance, which is what is often the issue in cases involving animal extrapolation, as when one wishes to know whether a chemical is a human carcinogen. A causal effect is more informative than a claim about positive causal relevance, and there are claims that fall between them in terms of specificity. This is illustrated by Donohue and Levitt's extrapolation, from European studies, of the claim that unwantedness doubles the chance of criminality later in life. In addition, one might wish to extrapolate a claim about causal structure, for instance, that socioeconomic status is a common cause of unwantedness and crime. The type of claim at issue matters because more stringent background assumptions are typically required for extrapolating more informative claims, a point which will be elaborated more fully in the subsequent section.

The definition itself does not assume that the causal structures in the model and target are represented by DAGs (with confounding arcs added) or that

the probability distributions satisfy the Markov condition with their respective causal structures. So, Definition 10.1 could be used for cases involving cyclic causal structures. However, for the purposes of this essay, I will assume that causal structures are acyclic and that the Markov condition is in force. Finally, $R(\Pi^*)$ is identifiable given $R(\Pi)$, P, P^*, and D means in effect that it is a logical consequence of these premises together with the Markov condition (see Pearl 2000, p. 77). It may seem surprising that extrapolation, which seems to be a type of inductive inference, would be defined in terms of logical consequence (of $R(\Pi^*)$ from $R(\Pi)$, P, P^*, and D). However, inductive inferences would be inevitably be involved in learning $R(\Pi)$, P, P^*, and D, so Definition 10.1 is not a covert expression of a deductivist perspective on scientific methodology. Moreover, stating the definition in this manner has the advantage of allowing for proofs about conditions in which extrapolation is and is not possible.

Definition 10.1 is similar to Pearl and Bareinboim's (2011, p. 8) definition of transportability. The main difference is that in Definition 10.1 here, extrapolation is premised on $R(\Pi)$, P, P^*, and D, while in Pearl and Bareinboim's definition of transportability, it is premised on $R(\Pi)$, P, P^*, G, and G^*, where G and G^* are the causal graphs for populations Π and Π^*, respectively. That is, Pearl and Bareinboim's definition is designed for cases in which the both causal structure and probability distribution are known for the target prior to the extrapolation. This is a rather restrictive assumption, as it entails that Pearl and Bareinboim's definition would not be useful for cases in which the causal structure in the target is not fully known. For example, animal extrapolation in toxicology often occurs in a background in which there is substantial uncertainty as to what adverse effect, if any, the chemical has in humans. In such cases, causal structure is part of what one wishes to learn by the extrapolation. In contrast, premising extrapolation on the selection diagram does not presume that the causal structure in the target is fully known, since a selection diagram indicates uncertainties about causal relations in the target population. Moreover, the proofs of the main theorems in Pearl and Bareinboim (2011) depend on knowing the selection diagram, not the causal structure of the target.

3.2 Making Adjustments

Since the model and target typically differ in some causally relevant respects, extrapolation usually requires making some adjustments. In other words, when extrapolation is *not* direct, some adjustment must be made to $R(\Pi)$ in order to infer $R(\Pi^*)$. Pearl and Bareinboim (2011) prove several theorems about how this can be done. The most general of these is their Theorem 3, which I restate below (but with "can be extrapolated" substituted for "is transportable" for consistency with the foregoing section):

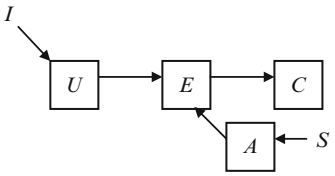


Fig. 10.3 The *U*-manipulated version of the selection diagram in Fig. 10.2

Theorem 3. *Let D be the selection diagram characterizing two populations, Π and Π*, and S the set of selection variables in D. The causal effect* $R = P^*(y|do(x))$ *can be extrapolated from Π to Π* if and only if the expression* $P(y \mid do(x), S)$ *is reducible, using the rules of do-calculus, to an expression in which no do-operator is conjoined with S*. (Pearl and Bareinboim 2011, p. 32)

As noted above, the causal effect of X upon Y, represented by the formula $P(y|do(x))$, is the probability distribution of Y conditional on an intervention on X. An intervention or manipulation on X is an exogenous cause that targets X alone and eliminates all other causes of X. To illustrate, consider the U-manipulated version of the selection diagram from Fig. 10.2 (i.e., the version of that diagram resulting from an intervention on the variable U). As shown in Fig. 10.3, this modified selection diagram eliminates the confounding arcs and adds a new variable labeled I, representing the intervention, with an arrow pointing directly into U. The do-calculus consists of three basic rules, derived from the Markov condition, for manipulating probabilities that contain *do*-operators, such as $do(x)$ (Pearl 2000, pp. 85–86).

To understand Theorem 3, first recall that the selection variables are assumed to account for all differences between the probability distributions in the model and target, so that $P^*(\bullet) = P(\bullet|s)$. Thus, any P-formula from which S is eliminable can be directly extrapolated from the model to the target, since in that case $P^*(\bullet) = P(\bullet|s) = P(\bullet)$. On the other hand, since $P(\bullet|s) = P^*(\bullet)$, any P-formula without a *do*-operator can be derived from the probability distribution of the target. In contrast, since no experimental manipulation was performed in the target population and selection variables are unmeasured, P probabilities with both an irreducible $do(u)$ and s are unidentifiable. Consider, then, how Theorem 3 applies to the selection diagram in Fig. 10.2, when $P(c|do(u))$ is the causal effect to be extrapolated. This proceeds as follows:

$$\begin{aligned} P^*(c|do(u)) &= P(c|do(u), s) \\ &= \sum_a P(c|do(u), a, s)P(a|do(u), s) \\ &= \sum_a P(c|do(u), a)P(a|do(u), s) \\ &= \sum_a P(c|do(u), a)P(a|s) \\ &= \sum_a P(c|do(u), a)P^*(a) \end{aligned}$$

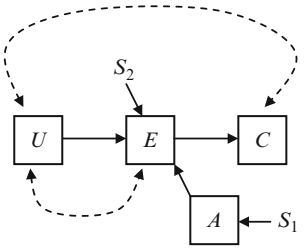


Fig. 10.4 A selection diagram in which the causal effect cannot be extrapolated

Using the rules of probability, the second equation expands $P(c|do(u), s)$ by summing over A. The next equation eliminates the s from $P(c|do(u), a, s)$, which is justified by the Markov condition applied to the selection diagram. Now all that needs to be done is to reduce $P(a|do(u), s)$ to probabilities in which $do(u)$ is absent, which is easily done in this case as A is independent of U.[9] The final equation results from a reapplication of the assumption that the selection variables account for all differences between model and target. The right-hand side of the final equation, then, consists of one probability, $P(c|do(u), a)$, that can be directly extrapolated from model to target and another, $P^*(a)$ that can be estimated using observational data sampled from the target population. The right-hand side of the final equation is an example of what Pearl and Bareinboim call a *transport formula*. A transport formula specifies how a causal effect in the model can be adjusted so as to be extrapolated to the target. In this example, then, $P^*(a)$ is the only probability in the transport formula that need be measured in the target.

However, whether extrapolation is possible depends on the selection diagram. For example, consider the selection diagram resulting from adding a selection variable pointing directly into E (as in Fig. 10.4). The presence of this selection variable blocks the step from the second to third equations in the reasoning above, because A does not d-separate S_2 from C. Indeed, $P(c|do(u))$ cannot be extrapolated from model to target given the selection diagram in Fig. 10.4 (see Appendix 2 for a proof of this claim). Yet the harmful psychological effects represented by E would plausibly be impacted by a variety of social, cultural, and economic factors that are likely to vary from one place and time to another. As a result, it would be difficult to justify the assumption that A or any other set of measured variables mediates all selection variables relevant to E. In general, then, extrapolating a causal effect is often very sensitive to difficult-to-justify assumptions about the absence of selection variables at crucial junctures in the selection diagram.

One way to deal with the problem of sensitivity to uncertain assumptions about the selection diagram is to be less ambitious about what one wishes to extrapolate. Causal effects are not the only type of causal claim that one might be interested in, and extrapolating other sorts of claims may require less demanding assumptions (see Steel 2008, chapters 5 and 6). Examples include claims about positive or

[9] In the selection diagram in Fig. 10.3, U is d-separated from A by the empty set, so $P^*(a \mid do(u)) = P^*(a)$ by rule 3 of Pearl's do-calculus.

negative causal relevance and claims about causal structure. For instance, suppose our concern was to extrapolate the claim that there is a causal chain from U to C. In this situation, we would not need to suppose that A mediates all selection variables impacting the effect of U upon C, nor would we have to assume that all causal paths from U to C pass through E. Extrapolating the claim that U has an effect on C only requires the premise that no circumstances are present in the target population that could completely eliminate this effect. Although it is easy to think of factors that could modulate the effect of U upon E, plausible circumstances that obliterate this effect entirely, if it exists, are much more difficult to come by. So, directly extrapolating a causal chain from U to C may be reasonable in this case. Such reasoning could be naturally extended into an extrapolation of positive causal relevance. The takeaway point of this example is that extrapolating claims about causal structure or positive causal relevance depend on much less stringent assumptions about the selection diagram. That makes such extrapolations more robust, though less informative.

This section has illustrated two central points concerning extrapolation. First, it is not necessary that the causal relationship to be extrapolated is the same in the model as in the target. Given knowledge of the probability distributions for the model and target along with the selection diagram, it can be possible to make adjustments to account for differences. Secondly, the conditions needed for extrapolation vary with the type of claim to extrapolated. In general, the more informative the causal claim, the more stringent the background assumptions needed to justify its transfer. This second point is very important for explaining how extrapolation can remain possible even when substantial uncertainty exists about the selection diagram. In the next, section I consider, in relation to the Donohue and Levitt study, how distinct levels of analysis can be helpful for assessing assumptions about the similarity of model and target.

4 Levels and Evidence

Extrapolation depends on background knowledge about ways in which the model and target are and are not likely to differ, knowledge that can be represented by a selection diagram. One obvious question, therefore, is where this knowledge comes from. Some similarities might be known only as a result of studies performed separately on the two populations. In other cases, the assumed similarity may be grounded in the acceptance of a common fundamental mechanism concerning, say, human psychology. For instance, it is natural to suppose that negative psychological impacts on a child of, say, insensitive and unconcerned parents are likely to be fairly stable across populations. In this type of situation, one might infer a causal relationship in the target on the grounds that it is found in the model and that model and target are unlikely to differ in that respect. However, it would be difficult to justify extrapolating a quantitative causal claim on the basis of such general theoretical considerations. Such background psychological knowledge might, for

instance, support extrapolating the claim that being born unwanted increases the risk of criminality, but it would be unlikely to justify the claim that unwantedness doubles that risk. It would be desirable, therefore, to have some means of testing assumptions about similarities between model and target. In this section, I consider how distinct levels of analysis can be helpful here.

Recall the connection between the concepts of extrapolation and integration as defined in Sect. 3.1. In extrapolation, a causal relationship R in the model is used, possibly with some adjustments, as a basis for inferring R in the target. Integration, by contrast, involves the extrapolation of R as part of a larger inference whose object is to infer another causal relationship, R'. Donohue and Levitt's reasoning fits this pattern because it extrapolates a claim about the effects of unwantedness on criminal convictions in order to draw an inference about the impact of abortion legalization in 1973 on crime rates in the 1990s. These observations suggest an approach for testing assumptions that underlie an extrapolation. Suppose that the causal relation R is directly extrapolated from model to target. Suppose, moreover, that R together with other background knowledge entails a further causal relationship R'. Then tests of R' will be indirect tests of the correctness of the direct extrapolation of R. In the Donohue and Levitt study, R is the claim that being born unwanted doubles the chance of criminal conviction later in life, while R' is the result of the scale-up model (or "back-of-the-envelope" calculations) described in Sect. 2. The results of the scale-up model, then, can be compared to estimates of the effect of abortion from Donohue and Levitt's state-level comparisons concerning abortion and crime rates. Donohue and Levitt characterize their statistical estimates of the impact of legalized abortion on crime as "roughly consistent, but somewhat larger than" their back-of-the-envelope result (2001, p. 391, p. 405). This rough consistency, then, is presumably taken as a reason for thinking that the scale-up model—including the extrapolated claim that being born unwanted doubles the chance of criminal conviction later in life—is a decent first approximation.

This example illustrates how differing levels of analysis can provide a means for testing assumptions about similarity and difference between model and target. Extrapolation at the level of a mechanism can be integrated with other information to generate an estimate of a macro-level causal effect, which then can in turn be compared with estimates directly made on the basis of macro-level data. The result of this process is an inference in which distinct lines of evidence, each with its own inevitable uncertainties, may mutually support or conflict with one another. The effect in the case of mutual support is, naturally, to strengthen the overall inference. Let us briefly consider the uncertainties in the present example. The uncertainties in the scale-up model are fairly easy to see. First, extrapolations rest on background assumptions concerning similarities and differences between model and target, assumptions which are often difficult to directly test. For instance, it is plausible that being born unwanted approximately doubles the chance of criminal conviction later in life in the USA just as found in the European studies described in Sect. 1. But it would be difficult to decisively eliminate the possibility that some divergence between the two populations exists that undermines this assumption, especially as

the key factor of being born unwanted is unmeasured in USA data. Similarly, many uncertainties would be involved in the details of the scale-up model through which this extrapolated result is integrated with other information to produce the estimate of the impact of legalized abortion in the 1970s on crime rates in the 1990s. Estimates of the impact of legalized abortion on crime based on national-level statistical data also confront a variety of uncertainties, for instance, concerning the proper modeling approach and how to account for other factors—such as the crack epidemic of the late 1980s—that might affect the results. Indeed, the discussions between Donohue and Levitt and their critics have predominantly focused on such issues (see Foote and Goetz 2008; Donohue and Levitt 2004, 2008; Joyce 2003). However, I should emphasize that the point here is definitely not to insist upon the infirmity of causal inferences grounded in extrapolation and observational data. Uncertainties frequently arise in experiments too, especially those involving human subjects (for instance, due to noncompliance, i.e., the failure of some subjects in the experiment to follow the experimental protocol). Such uncertainties are inherent in any attempts to learn about causation in large complex systems wherein numerous practical and ethical concerns restrict the types of studies that are possible. Consequently, scientific inference in such situations usually must build a cumulative case from a variety of lines of evidence none of which is decisive in isolation.

Although that may seem a rather obvious point, it does seem to get overlooked in some critical discussions of extrapolation. For instance, LaFollette and Shanks (1996) argue that results from animal experiments can never be extrapolated across species boundaries (e.g., from rats to humans) because causally relevant differences between populations are always present. The discussion of extrapolation in Sect. 3.2 has already illustrated several shortcomings with this line of argument. Extrapolation need not be direct, and it may be possible to adjust for relevant differences between the model and target. Moreover, some types of causal claims—such as claims about positive causal relevance—can be directly extrapolated even when considerable differences exist. Nevertheless, it is true that extrapolation is often haunted by the possibility that relevant differences between model and target have not been adequately accounted for. But this is only to say that there is often an unavoidable element of uncertainty inherent in extrapolations, just as there is in any other method for learning about causation in very complex systems. That in no way precludes extrapolations from being one useful line of evidence among others.

However, one might object that extrapolation can never be more than very weak evidence, useful only when information concerning the target population is grossly incomplete. Suppose that in initial stages of the investigation, studies performed on the target alone provide only rather uncertain evidence for the causal relationship and that in this context the extrapolation strengthens the overall case. Moreover, suppose that subsequent studies of the target population are able to make a compelling argument for causal claim in a way that does not require any reliance on the model.[10] This type of situation shows that the importance of extrapolation in

[10] Reiss (2010) suggests that an example discussed in (Steel 2008, chapter 5) follows this plot line.

providing evidence for a causal claim may wane as researchers become better able to study the target population. But there is a simple reason why such a course of events is unlikely to occur in relation to the impact in the USA of legalized abortion in the 1970s upon crime rates in the 1990s, namely, that being born from an unwanted pregnancy is unmeasured in the USA data. This variable can only be accurately measured in rather unusual circumstances, as illustrated by the European studies described in Sect. 1. Moreover, the claim that being born unwanted increases the likelihood of criminal activity later in life is a basic premise of the mechanism underlying Donohue and Levitt's hypothesis.[11] Furthermore, that mechanism plays an important role in reinforcing their arguments that the inverse correlation they find between abortion rates and lagged crime rates reflects a causal impact rather than the presence of some latent confounding factor. Finally, this important and apparently inescapable role of extrapolation does not, in and of itself, demonstrate any grave infirmity in Donohue and Levitt's overall argument. It is commonplace for mechanisms to play an important role in causal inference in social science. And in this case the extrapolation appears sufficient for the case at hand. First, a plausible case can be made for extrapolating claim about positive causal relevance (i.e., that being born unwanted makes a person more likely to be convicted of crimes later in life). Secondly, Donohue and Levitt's quantitative extrapolation—that being born unwanted doubles the chance of criminal conviction—need only be roughly accurate for the purposes of their argument, and this rough accuracy is supported by the compatibility of the results of their scale-up model and their estimates from national-level statistical data. Of course, the purpose of this chapter is not to defend the correctness of Donohue and Levitt's hypothesis. The point here is merely that extrapolation plays an important role in that argument and, furthermore, that this role of extrapolation is not a reason for thinking that they have failed to make a strong case for their conclusion. If Donohue and Levitt's statistical arguments are basically correct, then the extrapolation is one significant supporting plank in the overall structure of a strong argument. Therefore, this case belies the objection that extrapolation can be relevant as evidenced only in a context of massive uncertainty.

5 Conclusions

Let us recap the three interconnected philosophical themes relating to mechanisms and extrapolation that are highlighted by the case study discussed here. The first of these is that there are different types of causal claim that one might wish to extrapolate and that extrapolations of more informative causal claims typically

[11] For example, Levitt characterizes the hypothesis as resting on two premises: "(1) unwanted children are more likely to commit crime, and (2) legalized abortion leads to a reduction in the number of unwanted births" (2004, pp. 181–182).

rely on harder-to-justify assumptions. This point was illustrated by the Donohue and Levitt study, wherein extrapolating a qualitative claim about positive causal relevance rested on much firmer ground than extrapolating a quantitative claim about the strength of that impact. The second theme had to do with the connection between extrapolation and distinct levels of inquiry. Studies of causal relationships in social systems can focus on mechanisms linking individual people or take a bird's-eye statistical view of the population as a whole. Since the causal processes at these levels are not independent, claims about the one can have implications for the other. This idea is illustrated by the role of Donohue and Levitt's scale-up model in linking an extrapolated claim that being born unwanted doubles the chance of criminal conviction to statistical estimates of the impact of legalized abortion on crime. The correspondence of the results from these two lines of reasoning provides indirect support for the adequacy of that extrapolation as a rough approximation. The interplay between levels of inquiry leads to the third philosophical theme that extrapolation is normally one interwoven component of a complex and interdependent collection of arguments and, hence, is rarely a knockdown proof in its own right. Consequently, critiques which observe that extrapolations rarely if ever constitute definitive evidence sail wide of the mark. Building a case based on the coherence of multiple lines of imperfect evidence is the norm for social science and other sciences that study complex systems that are widely diffused across space and time. To insist otherwise is to misconstrue the nature of science and to obstruct applications of scientific knowledge to many pressing real-world problems.

Appendices

Appendix 1: Definition of d-separation

For completeness, I include the definition of d-separation, cited from Pearl (2000, pp. 16–17).

> *A path p is said to be d-*separated *(or* blocked*) by a set of nodes Z if and only if*
>
> 1. *p contains a chain* $i \rightarrow m \rightarrow j$ *or a fork* $i \leftarrow m \rightarrow j$ *such that the middle node m is in Z, or*
> 2. *p contains an inverted fork (or* collider*)* $i \rightarrow m \leftarrow j$ *such that the middle node m is not in Z and such that no descendant of m is in Z.*
>
> *A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.*

D-separation is important because it indicates all and only those probabilistic independence relationships entailed by the Markov condition. That is, the Markov conditions entails that X and Y are probabilistically independent conditional on Z in a DAG G if and only if Z d-separates X and Y in G.

Appendix 2: Causal Effect Cannot Be Extrapolated in Fig. 10.4

The selection diagram in Fig. 10.4 adds a selection variable pointing directly into E and thereby prevents deriving a transport formula by summing over A. Since $P(c|do(u))$ cannot be directly extrapolated, Pearl and Bareinboim's Theorem 3 tells us that $P(c|do(u))$ can be extrapolated given the selection diagram in Fig. 10.4 only if one of the following can be reduced by means of the *do*-calculus to a formula in which a *do*-operator and an s never occur in the same probability:

1. $\sum_{e} P(c|do(u), e, s)P(e|do(u), s)$
2. $\sum_{e,a} P(c|do(u), e, a, s)P(e, a|do(u), s)$

In 1, it is not possible to reduce $P(e|do(u), s)$ in this manner due to the selection variable pointing directly into E and the confounding arc linking U and E. The same reasons preclude reducing $P(e, a|do(u), s)$ in 2.

References

Dagg, Paul. 1991. The psychological sequelae of therapeutic abortion–denied and completed. *The American Journal of Psychiatry* 148(5): 578–585.

David, Henry, Zdenek Dytrych, Zdenek Matejcek, and Vratislav Schuller. 1988. *Born unwanted: Developmental effects of denied abortion*. New York: Springer.

Donohue, John, and Steven Levitt. 2001. The impact of legalized abortion on crime. *Quarterly Journal of Economics* 116(2): 379–420.

Donohue, John, and Steven Levitt. 2004. Further evidence that legalized abortion lowered crime: A reply to Joyce. *Journal of Human Resources* 39: 29–49.

Donohue, John, and Steven Levitt. 2008. Measurement error, legalized abortion, and the decline in crime: A response to Foote and Goetz. *Quarterly Journal of Economics* 123(1): 425–440.

Dytrych, Zdenek, Zdenek Matejcek, and Vratislav Schuller. 1988. In *The Prague cohort: Adolescence and early adulthood*, ed. H.P. David et al., 87–102. New York: Springer.

Elster, Jon. 1989. *Nuts and bolts for social science*. Cambridge: Cambridge University Press.

Foote, Christopher, and Christopher Goetz. 2008. The impact of legalized abortion on crime: Comment. *Quarterly Journal of Economics* 123(1): 407–423.

Fryer, Roland, Paul Heaton, Steven Levitt, and Kevin Murphy. 2005. Measuring the impact of crack cocaine. NBER Working Paper Series, no. w11318. Cambridge, MA: National Bureau of Economic Research.

George, Alexander, and Andrew Bennett. 2005. *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Gruber, Jonathan, Phillip Levine, and Douglas Staiger. 1999. Abortion legalization and child living circumstances: Who is the "marginal child"? *Quarterly Journal of Economics* 114(1): 263–291.

Hedstrom, Peter, and Richard Swedberg (eds.). 1999. *Social mechanisms: An analytical approach to social theory*. Cambridge, UK: Cambridge University Press.

Joyce, Ted. 2003. Did legalized abortion lower crime? *Journal of Human Resources* 39(1): 1–28.

Kincaid, Harold. 1996. *Philosophical foundations of the social sciences: Analyzing controversies in social research*. Cambridge: Cambridge University Press.

Lafollette, Hugh, and Niall Shanks. 1996. *Brute science: Dilemmas of animal experimentation*. New York: Routledge.

Levine, Phillip, Douglas Staiger, Thomas Kane, and David Zimmerman. 1999. *Roe v. Wade* and American fertility. *American Journal of Public Health* 89(2): 199–203.

Levitt, Steven. 2004. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic Perspectives* 18(1): 163–190.

Levitt, Steven. 2005. Abortion and crime: Who should you believe? *Freakonomics Blog*. http://www.freakonomics.com/2005/05/15/abortion-and-crime-who-should-you-believe/

Little, Daniel. 1992. *Varieties of social explanation*. Boulder: Westview Press.

Little, Daniel. 1998. *Microfoundations, method, and causation*. New Brunswick: Transaction Publishers.

Matejcek, Zdenek, Zdenek Dytrych, and Vratislav Schuller. 1988. In *The Prague cohort through age nine*, ed. H.P. David et al., 53–86. New York: Springer.

Pearl, Judea. 2000. *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearl, Judea, and Daniel Bareinboim. 2011. Transportability across studies: A formal approach. UCLA Cognitive Systems Laboratory, Technical Report (R-372), San Francisco.

Rasanen, Pirkko, Helina Hakko, Matti Isohanni, Sheilagh Hodgkins, Marjo-Riita Jarvelin, and Jari Tiihonen. 1999. Maternal smoking during pregnancy and risk of criminal behavior among adult male offspring in the northern Finland 1966 birth cohort. *The American Journal of Psychiatry* 156(6): 857–862.

Reiss, Julian. 2007. Do we need mechanisms in the social sciences? *Philosophy of the Social Sciences* 37(2): 163–184.

Reiss, Julian. 2010. Review: Across the boundaries: Extrapolation in biology and social science. *Economics and Philosophy* 26: 382–390.

Reyes, Jessica. 2007. Environmental policy as social policy? The impact of childhood lead exposure on crime. *The B.E. Journal of Economic Analysis & Policy* 7(1): Article 51.

Steel, Daniel. 2004. Social mechanisms and causal inference. *Philosophy of the Social Sciences* 34 (1): 55–78.

Steel, Daniel. 2008. *Across the boundaries: Extrapolation in biology and social science*. New York: Oxford University Press.

Steel, Daniel. 2011. Causality, causal models, and mechanisms. In *The Sage handbook on the philosophy of social science*, ed. Ian C. Jarvie and Jesús Zamora-Bonilla, 288–304. London: Sage.

Truman, Jennifer, and Michael Rand. 2010. Criminal victimization, 2009. *Bureau of Justice Statistics Bulletin* (NCJ 231327). Washington, DC: United States Department of Justice, Bureau of Justice Statistics.

Wadsworth, Tim. 2010. Is immigration responsible for the crime drop? An assessment of the influence of immigration on changes in violent crime between 1990 and 2000. *Social Science Quarterly* 91(2): 531–553.

Chapter 11
Causality, Impartiality and Evidence-Based Policy

David Teira and Julian Reiss

Abstract The overall aims of this chapter are to compare the use of randomised evaluations in medicine and economics and to assess their ability to provide impartial evidence about causal claims. We will argue that there are no good reasons to regard randomisation as a sine qua non for good evidential practice in either science. However, in medicine, but not in development economics, randomisation can provide impartiality from the point of view of regulatory agencies. The intuition is that if the available evidence leaves room for uncertainty about the effects of an intervention (such as a new drug), a regulator should make sure that such uncertainty cannot be exploited by some party's private interest. We will argue that randomisation plays an important role in this context. By contrast, in the field evaluations that have recently become popular in development economics, subjects have incentives to act strategically against the research protocol which undermines their use as neutral arbiter between conflicting parties.

List of Abbreviations

FDA	Food and Drug Administration
NGO	Non-Governmental Organisation
RCT	Randomised Clinical Trial
RFT	Randomised Field Trials

D. Teira (✉)
Departamento de Lógica, Historia y Filosofía de la ciencia, UNED, Paseo de Senda del rey 7, 28040 Madrid, Spain
e-mail: dteira@fsof.uned.es

J. Reiss
Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK
e-mail: julian.reiss@durham.ac.uk

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_11,

1 Introduction

Randomisation, the assignment of experimental subjects to treatment groups by means of a random number generator, was first systematically applied in psychic research in the late nineteenth century and became popular in statistics after Ronald Fisher advocated its use in 1926 (Hacking 1988). In medicine and development economics, the two sciences we will focus on in this chapter, randomised trials are now widely regarded as the 'gold standard' of evidence. The overall aims of this chapter are to compare the use of randomised evaluations in these two sciences and to assess their ability to provide impartial evidence about causal claims. In short, we will argue that there are no good reasons to regard randomisation as a sine qua non for good evidential practice in either science. However, in medicine, but not in development economics, randomisation can provide impartiality from the point of view of regulatory agencies. The intuition is that if the available evidence leaves room for uncertainty about the effects of an intervention (such as a new drug), a regulator should make sure that such uncertainty cannot be exploited by some party's private interest. We will argue that randomisation plays an important role in this context. By contrast, in the field evaluations that have recently become popular in development economics, subjects have incentives to act strategically against the research protocol which undermines their use as neutral arbiter between conflicting parties.

2 Background: Randomised Clinical Trials as a Public Policy Tool

Randomised clinical trials (RCTs) are medical experiments in which alternative treatments for a condition are administered to at least two groups of patients in order to see which one is the safest and most effective for future cases. Unlike other experimental designs in medicine and elsewhere, RCTs have achieved some public notoriety throughout the last five decades thanks to the role they play in pharmaceutical regulation. The commercial distribution of novel drugs will only be authorised by regulatory agencies such as the American Food and Drug Administration (FDA) if their safety and efficacy is proved in two RCTs.

RCTs come to solve a problem in public policy: by their own means, consumers cannot ascertain the quality of a drug, either by simple inspection of their appearance (shape, size, smell, etc.) or by their price. Depending on the circumstances of the patient, the natural rate of variability of their effects (positive or negative) prevents a reliable assessment on the basis of individual experience alone. When buyers or sellers cannot directly determine the quality of a good or service, economic *adverse selection* can lead to the elimination of all trade in a market

(Akerlof 1970; Wilson 2008): putting it very simply, if consumers know that there is a percentage of cheap, bad-quality drugs, they will be reluctant to pay the price requested for good-quality compounds, and the producers of the latter may end up leaving the market. Regulation may be justified to remedy this type of market failure whose consequences can be fatal for the consumers (because they need the good-quality drugs).

RCTs are one way to provide the grounds for an evidence-based pharmaceutical policy: the regulator will make a decision on the marketing of a new drug depending on the evidence RCTs yield about its safety and efficacy. One may wonder, however, why RCTs are regarded as credible, given the conflicts of interest that pervade the pharmaceutical markets. Historically, physicians, pharmacists and patients have supported their favourite treatments, seeking whatever evidence confirmed their views and questioning, with the same passion, the quality of any piece of adverse evidence. Why should they now accept RCTs?

A standard sociological response is because in democratic societies RCTs provide an appearance of *mechanical objectivity* that seems more acceptable than mere expert clinical judgement: the statistical apparatus underlying RCTs proceeds impartially, impervious to the particular interests that may bias the judgement of the individual expert. But, so the standard response continues, mechanical objectivity is a mere appearance caused by numbers whose statistical justification lay audiences cannot grasp (see, for instance, Porter 1995; Marks 1997).

In these sociological accounts 'mechanical objectivity' contrasts with 'expert judgement'. The so-called evidence-based medicine (e.g. Sackett et al. 1996) is a paradigmatic example for the perennial attempts to replace the latter by the former. In pre-evidence-based medicine, the standard approach to assessing the efficacy of new treatments was heavily influenced by clinicians' judgements. But clinicians, like all experts, may be inattentive, ill informed, partial (to this or that therapy) or otherwise biased. Moreover, an expert's decision is not transparent to outsiders (in this case, patients). There are therefore good reasons to limit the influence of the clinician's judgement to a minimum and replace it with 'objective evidence'. Objective evidence is sometimes called 'mechanical' when it is produced by mechanical methods such as RCTs. An RCT is a mechanical method in that its implementation follows strict and explicit rules – divide the test population into two groups by means of a random allocation mechanism, blind subjects and treatment administrators, follow specific stopping rules, etc. Unlike expert judgements, such mechanical rules are transparent. This means that they can be publicly debated, scrutinised and criticised.

Over the last 10 years, philosophers of science such as Nancy Cartwright and John Worrall have challenged the epistemic foundations for RCTs (e.g. Cartwright 2007; Cartwright and Munro 2010; Worrall 2002, 2007). They appraise RCTs as tools for causal inference. In their – philosophers' – approach, impartiality is at best a by-product of causal analysis: if one can establish objectively that a drug is effective in curing a given condition, this judgement is independent from whatever

interest there might be at stake in the experiment. Both Cartwright and Worrall conclude that RCTs are not completely reliable tools for causal inference, and therefore, we should consider alternative sources of evidence for our regulatory decisions, including expert judgement. In the following section, we will examine Cartwright's criticism in more detail.

3 RCTs as 'Gold Standard' of Evidence for Causal Claims

According to Nancy Cartwright, RCTs are just one method among others for warranting causal claims. In her terminology, it is a 'clinching' method in that it proceeds in a deductive fashion: *if* its assumptions are met and the observable evidence is positive, *then* we can safely affirm the causal claim. However, the premises are restrictive, and therefore, the range of conclusions narrow (Cartwright 2007).

Ideal RCTs test causal claims about the narrow efficacy claims of drugs in a given population following Mill's method of difference. Given an observed outcome (O), we study the probability of the difference between outcomes with and without the treatment intervention (T) in two groups of patients drawn from a population φ. In these two groups, all causally relevant factors other than T are equally distributed. (This is what randomisation is supposed to achieve; more on that later.) Therefore, the observed difference in O must be an effect of T. To show that the effectiveness claim follows deductively, we need a number of further assumptions.

The first assumption is a *causal fixing condition* (Cartwright and Munro 2010, p. 261): the probability of an effect is fixed by the values taken by a full set of its causes. Cartwright adopts a version of Patrick Suppes' probabilistic theory of causality which states that for an event-type T preceding event-type O in a population Θ,

$$\begin{aligned} &\text{T causes O in } \Theta \text{ iff } \; P(O/T\&K_i) > P(O/\neg T\&K_i) \\ &\qquad \text{for some subpopulation } K_i, \text{ with } P(K_i) > 0. \end{aligned}$$

Cartwright further assumes that the individuals in the sample are all governed by the same causal structure CS, described by a probability distribution P. According to Cartwright, 'P is defined over an event space $\{O, T, K_1, K_2, \ldots, K_n\}$, where each K_i is a state description over 'all other' causes of O except T'. Conditioning on these potential *confounding factors*, we can attribute the remaining difference between $P(O/T\&K_i)$ and $P(O/\neg T\&K_i)$ to the causal link between T and O. In an ideal RCT, claims Cartwright (2007, p. 15), the K_i are distributed identically between the

treatment and control groups. Hence, any difference in outcome between groups can be causally attributed to T in at least one K_i relative to the causal structure CS described by P. This is the conclusion ideal RCTs can clinch. However, according to Cartwright, we need further assumptions still if we want to generalise this conclusion to some target population Θ.

If we want to affirm, for instance, that T causes O in at least some members of Θ, Cartwright (2007, p. 17) argues, we need assumptions of this kind:

(a) At least one of the subpopulations (with its particular fixed arrangement of 'other' causal factors) in which T causes O in φ is a subpopulation of Θ.
(b) The causal structure and the probability measure are the same in that subpopulation of Θ as it is in that subpopulation φ.

The warrant for these assumptions too is supposed to come from randomisation, but we cannot judge whether a group of patients constitutes a random sample without a previous idea of what factors are to be equally represented (Cartwright 2007, p. 18). In a trial, we want to form, on the one hand, two treatment groups that are balanced with respect to known relevant prognostic factors. On the other hand, we want to avoid unknown confounders to affect the result. Randomisation supposedly helps us in achieving both goals, but it is neither necessary nor sufficient to that effect. By sheer chance, a random allocation may yield an unbalanced distribution of the prognostic factors between the treatment groups (these are called 'baseline imbalances'). This may bias the comparison between treatments and invalidate the experimental results, and when imbalances occur, trialists usually try to correct them (e.g. by repeating the randomisation). Unknown confounders may differentially influence the outcome in one of the groups even after a randomised allocation of treatments. Further randomisations at each step in the administration of the treatment (e.g. which nurse should administer the treatment today?) may avoid such interferences, but this is impracticable. Declaring such disturbances as negligible, as many experimenters do, lacks any justification in the assumed statistical methodology (Urbach 1985; Worrall 2007).

Both the correction of imbalanced allocations and the decision to randomise at different stages of the trial beyond the allocation of treatments require extra-statistical expert judgement. Against the ideal of mechanical objectivity, we need an expert who can handle different sources of evidence other than the trial to justify the acceptance of assumptions (a) and (b). More precisely, we need someone who can certify that randomisation, the main warrant of (a) and (b), has indeed worked. Without this judgement, subjective and intransparent as it may be, we cannot safely generalise the conclusions of the trial to its target population, i.e. ascertain its external validity. Expert judgements are naturally fallible too, but, according to Cartwright (2007, p. 19), to rely on mechanical methods without expertise and watch out for failures is no satisfactory response.

4 External Validity and Impartiality in Regulatory RCTs

Today, FDA is probably the institution that makes the most systematic use of RCTs for regulatory purposes in the world, but has not always been so. Between 1900 and 1950, expert clinical judgement was the main criterion in the assessment of the properties of pharmaceutical compounds in the United States as well as in other countries such as Britain. An experienced clinician would administer the drug to a series of patients he would consider likely to benefit. His or her conclusions would be presented as a case report, informing of the details of each patient's reaction to the treatment. The regulatory authorities in the United States and Britain arranged official drug testing depending on the standards adopted by the research community within their respective medical professions. Until the 1960s, regulatory decisions were fundamentally based on expert judgements of this sort. Clinical judgement came to be discredited in the United States because a group of methodologically minded pharmacologists took over the FDA and imposed their views on the superiority of RCTs through regulatory means. This was the triumph of mechanical objectivity against expert judgement.

During the 1960s and 1970s, RCTs became mandatory for regulatory decisions in different degrees. In the United States, before the 1960s, the FDA was entitled only to test the safety but not the efficacy of pharmaceutical compounds. In the late 1950s, there were voices in the FDA demanding stricter testing standards linking safety and efficacy, under increasing public mistrust in the pharmaceutical industry, prompted in part by the thalidomide scandal.

Under the trade name *Contergan*, one million West Germans consumed thalidomide as a sedative in the early 1960s and many more people around the world after that. Reports showing an association between the drug and peripheral neuropathy were soon published in medical journals. Later reports of serious birth defects when the drug was consumed by pregnant women surfaced. Only then did the manufacturer withdraw the drug from European markets. Eight thousand babies had been already born with severe deformities. At that point, there was no clear regulatory standard about the safety of a compound, neither in the United States nor in Europe.

The thalidomide scandal gave them the opportunity to put their views in effect in the 1962 Drug Efficacy Amendment to the Food, Drug and Cosmetics Act. It required from the applicant 'adequate and well-controlled clinical studies' for proof of efficacy and safety (although the definition of a well-controlled investigation would not be clarified until 1969, when it was formally quantified as two well-controlled clinical trials plus one prior trial or posterior confirmatory trial). It has been claimed, correctly in our view, that this set of regulations created the modern clinical trial industry (Carpenter and Moore 2007). In the following three decades, pharmaceutical funding would boost the conducting of RCTs (by the thousands) in the United States and abroad.

The regulatory approach of the FDA surely constitutes a canonical instance of an evidence-based policy or, more precisely, an RCT-based policy. It is worth noting

that the FDA does not take the external validity of an RCT for granted. A drug trial is usually divided in four phases. Phase I focuses on finding the appropriate dosage in a small group of healthy subjects (20–80 patients); toxicity and other pharmacological properties of the drug are examined. In phase II, between 100 and 200 hundred patients are closely monitored to verify the treatment effects. If treatment effects are detected, a third phase involving a substantial number of patients begins in which the drug is compared to the standard treatment. This is usually referred to as 'the' RCT. If the new drug proves to be at least as good as the existing therapies and the pharmaceutical authorities approve its commercial use, phase IV starts: the drug is freely prescribed and sold; adverse effects are monitored and morbidity and mortality studies are undertaken.

In other words, the FDA, as other regulatory agencies, does not take the external validity of the RCTs for granted when it approves a new substance. In the post-market surveillance phase IV, the FDA collects adverse event reports from various sources and conducts epidemiological studies to assess their relevance, keeping track of the validity of the results of their trials in the general population. The authority and resources of the FDA at this stage are disproportionately smaller than at any previous point in the approval process. And the assignment is large: apart from monitoring adverse reports, the agency also has to consider issues in labelling, advertising or the inspection of production and storage facilities, to name but a few. Hence, one should appraise the figures collected at this fourth phase *cum mica salis*. But we think they are relevant in the context of our analysis of the reliability of RCTs.

Changes in drug labelling constitute a first approximation to the number of minor or major failures at phase III trials. According to Dan Carpenter (2010, p. 212), the FDA has relied on these changes as a cheap regulatory strategy, given the available resources, as compared with pursuing withdrawal or a change in advertising and prescribing practices (advertising and prescription are only lightly regulated in the United States as compared to Europe). As long as the label records potential safety threats, the FDA can claim that the consumer has been warned. Each label change requires an application for approval, which creates a data record. Dan Carpenter has compiled it in the following table, where it is compared to other product changes for the same periods (Fig. 11.1):

Carpenter (2010, p. 623) summarises it as follows: from 1980 to 2000, the average new molecular entity received five labelling revisions after approval, about one for every 3 years of marketing after approval. Only one in four drugs had no labelling revisions at all. The data are obviously too coarse to decide what went wrong, if anything, in the phase III RCTs. Several explanations are possible: for instance, the trials might have been too brief to detect adverse effects (e.g. toxicity or cardiovascular events). In the context of Cartwright's analysis, we may suspect that the patients' sample might have been unrepresentative of the patients that finally used the therapy. In a rough sense, the list of prognostic factors and potential confounders used to define the eligibility criteria was incomplete and randomisation could not correct this flaw. If we take external validity in an equally rough sense, Carpenter's data would suggest that Cartwright points out correctly the

Drug Changes Requiring a Supplemental NDA, 1970–2006

	1970–1974	1975–1979	1980–1984	1985–1989	1990–1994	1995–1999	2000–
Chemistry Revisions	2	376	3,710	7,728	5,664	8,520	258
Manufacturing Revisions	0	492	910	1,045	1,063	2,229	1,936
Package Changes	38	465	757	733	573	847	994
New or Modified Indications	3	6	7	76	121	273	294
Control Supplements	242	2,516	3,710	2,138	1,902	2,885	4,357
Labeling Revisions (SLR)	529	1,968	2,005	2,360	1,909	2,341	4,472
Other Label Changes	0	0	168	1,998	3,588	2,923	1,672

Fig. 11.1 Drug changes requiring a supplemental new drug application, 1970–2006 (Carpenter 2010, p. 613)

limitations of causal inference in RCTs: the two phase III RCTs that granted the approval of the drug do not usually capture the full range of effects of a drug.

However, it is useful to compare these figures with drug withdrawals. We should always bear in mind that phase III trials are testing the safety and efficacy of a compound, but not their full range of effects, which are only seen in phase IV. The figures should be taken again with caution, since, as Carpenter (2010, ch. 9) warns, the negotiation of each withdrawal depends on a number of circumstances outside and inside the agency, among which a prominent one is the time constraints for the review process (cf. Carpenter 2010). However, very few compounds have been withdrawn from the market in the United States during the last five decades for lack of safety or efficacy after receiving the authorisation of the FDA: if we exclude the drugs approved just before the new legal deadline established in 1992, for which security issues seem to be more prominent, between 1993 and 2004 only 4 out of the 211 authorised drugs were withdrawn.

If we thus take label revisions and market withdrawals as rough indexes of the external validity of the regulatory trials approved by the FDA, we may conclude that the procedure is not foolproof (in the sense of anticipating every safety threat a drug may pose), but that it does not fare completely badly either. Its main effects are reasonably well anticipated. Of course, this is a black box argument: we know that the four-phase regulatory system at the FDA screens off dangerous compounds, but perhaps this is just because the pharmaceutical industry does not dare to submit any potentially dangerous new compound. Assuming that the FDA system works (and very few people question that it does), RCTs certainly do not explain its success

alone. Lots of formal and informal causal knowledge are acquired in the first two phases, and it is put to test not only in the RCTs, but by the subsequent epidemiological surveillance, if the drug is approved.

Expert judgement contributes to all the four phases, and even if the decision to authorise a drug is taken at the third phase, the decision is not *mechanical*, precisely because the external validity of a trial cannot be taken for granted. But this is something pharmacologists have known right from the beginning: the regulatory system at the FDA was established to deal with causal uncertainty. During most of the second half of the twentieth century, pharmaceutical research advanced through a so-called molecular lottery: compounds were synthesised and tested on animals without any clear theoretical guidance, and even when they had an interesting therapeutic effect, there often was no grasp of the precise mechanism responsible for it. RCTs allowed pharmacologists to deal with this causal uncertainty about drugs, focusing on the probability of attaining certain treatment outcomes under a given range of prognostic factors.

But if RCTs rely so crucially on expert judgement, we may wonder why they were considered an improvement over the older methodology. One of us has defended elsewhere that all the involved parties explicit sought a testing methodology with warrants of impartiality against the potential conflicts of interest arising in the trial (Teira 2011a, b). The 1950s saw a boom in industrial drug production (some of which were 'wonder drugs' such as antibiotics, but many were just combinations of already available compounds) and, simultaneously, in pharmaceutical advertising that caused much confusion among practitioners about the therapeutic merit of each product (Marks 2000, p. 346). For therapeutic reformers, RCTs with their strict research protocol provided the information about drugs that 'sleazy advertising' was trying to disguise with 'badly scissored quotes', 'pharmaceutical numbers racket', 'detail men' visits and so forth (Lasagna 1959, pp. 460–461). Adopting RCTs as a regulatory tool allowed the FDA to justify the impartiality of their decisions about treatments before patients, physicians and the pharmaceutical industry (Teira 2011a, b).

In this respect, randomisation was considered more as a debiasing procedure than a tool for causal inference. Randomisation prevents researchers from allocating treatments to patients according to their personal interests, so that the healthiest patients get the researcher's favourite therapy. As mentioned above, such unbalanced allocations can happen nonetheless by chance. But randomisation is still a warrant that the allocation was not done on purpose with a view to promoting somebody's interests. A priori, the experimental procedure is impartial with respect to the interests at stake.

Of course, in Cartwright's causal approach, randomisation would just be a tool for controlling the probabilistic dependences arising from selection biases in the experimental and the control group, making sure that they are the same in both treatment groups except for the treatment. To understand randomisation in this way is perfectly appropriate from a methodological point of view, but this is not how it was understood by those who introduced it into the regulatory system. Given how little was known about causation in RCTs at the time, randomisation was not sought

for its contribution to causal analysis but rather for preventing anyone to exploit this uncertainty about causation for her own benefit. The same justification applies to other standard features of RCTs such as masking treatments or having pre-established decision rules for the interpretation of the results (as in significance testing): they provide a priori warrants for the impartiality of the trial.

The research protocol in RCTs constrains expert judgement at various critical points in the generation and interpretation of clinical evidence during a trial. In this sense, pharmaceutical regulation is based as much on the *impartiality* of its evidential base as on the accuracy of its causal conclusions. Perhaps there are other sources of evidence whose external validity is as good as RCTs, but it is an open question if they can be as impartial as the latter. Impartiality is crucial for public policy, and it seems a defensible decision to adopt RCTs instead of mere expert judgement for regulatory purposes: at least, the former provide certain warrants of impartiality.

An obvious objection, of course, is that RCTs are not actually impartial. There is evidence showing, for instance, a *sponsor* bias: industry-funded trials published are more likely to support the experimental therapy than the standard alternative, despite their good methodological quality. We know that RCTs do not control every possible source of bias, e.g. the research protocol does not impose any constraint on the research question that trials should address, and there is no obvious way to decide which one should it be. But rather than an objection against RCTs, it should be a general caveat about every possible source of clinical evidence: the easier it is to manipulate the method, the less we should rely on the evidence it produces for regulatory purposes.

Summing up, despite the problems with their external validity, regulatory RCTs have been reasonably efficient in keeping the American pharmaceutical market clear of unsafe or inefficacious compounds. Moreover, despite all the label revisions, the American public has considered the FDA a reliable regulator (Carpenter 2010), and we contend that this is because RCTs provided a warrant of impartiality for their decision, despite the inherent uncertainty of phase III trials. If we had full information about the effects of a therapy, impartiality would be warranted by default. However, short of that, we need to make sure that a regulatory decision is fair despite their inherent imperfection.

5 The Impartiality of Randomised Field Evaluations

The assessment of public policy programmes through large-scale randomised field evaluations (RFEs) is already several decades old (the 1968 New Jersey negative income tax experiment is often considered to be a pioneering example). Usually the interventions assessed deal with one or another aspect of the welfare of large populations, and testing them is expensive, though the cost of the actual implementation of the programme would be significantly more so. Around 200 RFEs were

run in the United States between 1960 and 1995 (Orr 1999), with more or less convincing results.

In the last decade, there has been an explosion of interest in RFEs among development economists. Several programmes for improving health or education, different microfinance and governance schemes have been tested in a number of developing countries. A success story is the PROGRESA programme implemented in Mexico in 1998. PROGRESA aimed at improving school performance through a system of direct transfers conditional on family income, school attendance and preventive health measurements. The amount of the allocation, received directly by the mothers, was calculated to match the salary of a teenager. In order to test the effects of PROGRESA (and with a view to secure its continuation if there was a change in government), a team at the Ministry chose 506 villages, implementing PROGRESA in a randomly selected half of them. The data showed an increase in teenager enrolment in secondary education significantly higher in the experimental group, with concomitant improvements in the community health. The experiment was considered convincing enough to ground the extension of the scheme to more than 30 countries.

The boom of RFEs in development economics may owe something to their costs: in developing countries, the costs for running these programmes are significantly lower than, say, in the United States, and non-governmental organisations can implement them in a quick and efficient manner. But there is also a sense of political opportunity among these social experimentalists. A leading one, Esther Duflo, puts it as follows: just as RCTs brought about a revolution in medicine, RFEs can do the same for the assessment of our education and health policies in fighting poverty (Duflo 2010, p. 17).

Nonetheless, Duflo acknowledges that RFEs can involve many methodological pitfalls. Randomisation is a case in point. Field experimentalists in economics expect it to provide a solid foundation for causal analysis, and we have already discussed Cartwright's criticism of this idea. In this section we discuss further whether we can take RFEs in development economics to be impartial. More precisely, our question is whether randomisation is as credible warrant of impartiality in field trials development in economics as it is in medical RCTs. We think not.

Let us present our case by drawing on an analysis due to James Heckman. In 1992, Heckman published a seminal paper containing 'most of the standard objections' against randomised experiments in the social sciences. Heckman focused on the non-comparative evaluation of social policy programmes, where randomisation simply decided who would join them (without allocating the rest to a control group). Heckman claimed that even if randomisation allows the experimenters to reduce selection biases, it may produce a different bias. Specifically, experimental subjects might behave differently if joining the programme did not require 'a lottery'. Randomisation can thus interfere with the decision patterns (the causes of action) presupposed in the programme under evaluation.

Let us briefly present Heckman's semiformal analysis. Let D represent participation in a programme and Y the outcome of participating. These two variables are related as follows:

$$Y = Y_1 \quad \text{if D} = 1 \quad \text{[The outcome of participating]}$$

$$Y = Y_0 \quad \text{if D} = 0 \quad \text{[The outcome of not participating]}$$

Heckman presumes that the values of Y_0 and Y_1 are causally determined by some umbrella variables X_0 and X_1:

$$Y_1 = g_1(X_1)$$

$$Y_0 = g_0(X_0)$$

If we are evaluating a training programme, and Y_1 is the outcome attained by the participants, we may presume it to be determined by their previous education, age, etc. (X_1). Participation in the programme is determined in turn by another umbrella variable Z, with a subset of values Ψ:

$$\text{If Z} \in \Psi, \quad \text{D} = 1; \quad \text{otherwise,} \quad \text{D} = 0$$

For instance, participation may depend on certain values of income, employment, etc., all captured by Z. The collection of explanatory variables in the programme assessment is thus $C = (X_0, X_1, Z)$: the outcome depends on certain antecedent factors (captured by X_i) and on participation (Z). We usually do not have full information about C: the available information is represented by C_a. If we conduct an experiment to assess this programme, we try to determine the joint probability distribution of Y_1, Y_0, D conditional of a particular value of $C_\text{a} = c_\text{a}$:

$$F(y_0, y_1, d|c_\text{a}).$$

In order to make his first objection, Heckman suggests we should distinguish between regular participation in a programme (captured by D) and participation in the programme in an experimental regime, where participation is randomised. This is captured by a second variable D^*:

$$D^* = 1 \quad \text{if a person is at risk for randomisation.}$$

$$D^* = 0 \quad \text{otherwise.}$$

If p is the probability of being accepted in the programme after randomisation, the possibility of testing the programme through randomised tests depends on the following assumption:

$$\Pr(D = 1|c) = \Pr(D^* = 1|c, p).$$

In other words, we need to assume either that:

1. Randomisation does not influence participation, or
2. If it does influence participation, the effect is the same for all the potential participants, or
3. If the effect is different, it does not influence their decision to take part in the programme.

Heckman's main objection is that randomisation tends to eliminate risk-averse persons. This is only acceptable if risk aversion is an irrelevant trait for the outcome under investigation – i.e. it does not feature in C. However, even if irrelevant, it compels experimenters to deal with bigger pools of potential participants in order to meet the desired sample size, so the exclusion of risk-averse subjects does not disrupt recruitment. But bigger pools may affect in turn the quality of the experiment, if it implies higher costs. One way or another, argues Heckman, randomisation is not neutral regarding the results of the experiment.

Heckman's analysis is causal: randomisation can create a self-selection bias distorting the sample of participants on which any inference should rest. We are going to argue that it is impossible to correct this self-selection bias without putting in question the impartiality of the trial. The threat of partiality does not come in this case from the researchers but from the participants themselves. In RFEs, participants may have their own preferences about the compared treatments, and the risk aversion elicited by randomisation is just one of them: people may prefer to make choices about treatments. In order to preserve randomisation and to correct self-selection biases, Duflo and her coauthors try to control the participants' preferences by blinding, i.e. by disguising or hiding the randomised nature of the experiment. We argue that these attempts assume the indifference of the participants regarding the experimental outcome. If the participants have strong preferences about the outcome, masking randomisation will not be enough to enforce the experimental protocol.

According to Banerjee and Duflo (2009), we can avoid the self-selection bias if we either disguise or hide randomisation. Both solutions are feasible in many programmes, at least if we conduct the experiment in a developing country. As to the former, randomisation can be disguised as a lottery by which the scarce resources of the programme are allocated. If the potential participants perceive this lottery as fair, it may not dissuade them from taking part in it. The fairness of lotteries as allocating procedures can be certainly defended on theoretical grounds (Stone 2007), and we know that there is empirical evidence about the acceptability of unequal outcomes when they come from a lottery perceived as fair (Bolton et al. 2005). However, not everybody likes lotteries, even fair ones: for instance, surveys show that people oppose the use of lotteries by colleges and universities in order to choose which students are admitted (Carnevale et al. 2003).

It is an empirical question to be solved on a case-by-case basis if disguising randomisation as a lottery influences participation. Banerjee and Duflo certainly acknowledge that even fair lotteries can provoke a self-selection depending on the way they are presented: if the participants in the control group are told that the experimental treatment will be available to them in the future (once the resources are gathered), this may affect their willingness to participate or their compliance. In addition, organising a lottery to distribute aid seems to be politically controversial for governments that are expected to serve an entire population (Duflo et al. 2007, p. 21).

Hiding randomisation altogether from participants seems a more effective strategy. As Banerjee and Duflo observe, 'ethics committees typically grant an exemption from full disclosure until the endline survey is completed, at least when the fact of being studied in the control group does not present any risk to the subject' (2009, p. 20). Participants in the experimental group will not know how they got involved, and those in the control group may never know they have been excluded. If the latter live in different villages, as it often happens in trials run in developing countries, they may not get to know about the experimental treatment. In this way a totally different scenario arises: in order to avoid a self-selection bias, we deceive the participants about the comparative structure of the experiment. The experimenters are assuming here that participants only care taking part in a lottery, but, as a matter of fact, they may also have preferences about the treatments tested. They may want to get one rather than the other. Or, if they understand the nature of the experiment, they may even have a favourite treatment that they want to see succeed – e.g. we may well imagine parents preferring direct allocations of cash to send their kids to school rather than paid meals. If these preferences exist, disguising randomisation will only succeed to the extent that the disguise is successful: the participants have been 'blinded' to the comparison, but shall we just assume that such blinding is successful?

There is some evidence that deception in medical trials can fail. Patients have preferences about treatments, and they usually neither understand nor like randomisation (Featherstone and Donovan 2002): their compliance is usually explained by their lack of alternatives to get access to experimental treatments, they would not get the medication outside the trial. And they play by the research protocol only to a point: they try to find out which treatment they are receiving (and if they succeed, this has an effect on the experiment). However, in most medical trials, the researchers have means to make patients comply with the research protocol, e.g. they may mask the treatments well enough for an ordinary participant not to be able to distinguish them. They would need a laboratory. Whether they have access to a laboratory often depends on the social organisation of the patients. The testing of early anti-AIDS treatments in the USA, documented by Epstein (1996), illustrates this point: the participants wanted to have experimental treatments and not placebos, so they resorted to all sort of strategies to make sure they would receive the treatment, drawing on their connections in the gay activism networks. Many abstained from taking part in trials if they didn't think the drug was promising enough (in order to remain 'clean' and thus eligible for other tests); those who

participated exchanged the pills between them (at the cost of halving the dose) or took them to independent laboratories to verify the active principle. They completely undermined the trial protocol.

Drug trials in developing countries illustrate how access to experimental treatments becomes a politically contentious issue within the country (Macklin 2004; Petryna 2009). We can probably expect the same from RFEs in economics: if they address interventions about which the potential participants have preferences, randomisation may elicit a different type of self-selection. Participants may behave differently depending on their taste for a treatment, over-complying if they want it to succeed or the opposite if they want to see it fail. Randomisation will only succeed in breaking any correlation between the participants' preferences and the trial outcome if these former remain ignorant about the comparative nature of the experiment. But if they have strong preferences about the treatments, how far can we go in deceiving them about the comparison?

In order to control for such post-randomisation effects, Duflo et al. (2007) suggest two additional strategies. The first is continue collecting data after the experiment is terminated in order to verify whether the interaction with the experimenter was making any difference in the behaviour of the participants (e.g. Duflo and Hanna 2006). One way or another, we need participants to remain ignorant about the controls: they should not know they are still being observed. And we need to test this ignorance, just as in medical trials with blinding; we just cannot take it for granted.

To sum up, in RFEs, randomisation may generate a self-selection bias; we can only avoid with a partial or total masking of the allocation procedure. We have argued that this is a viable solution only insofar as the trial participants do not have strong preferences about the trial outcome. If they do, we cannot assume that blinded randomisation will be a control for their preferences unless we test for its success. We will only be able to claim that the trial has been impartial regarding the participants' preferences if we have a positive proof of them being ignorant of the comparative nature of the experiment. Hence, in RFEs, randomisation is not a strong warrant of impartiality per se: we need to prove in addition that it has been masked successfully.[1]

6 Can Field Trials Ground an Evidence-Based Policy?

In order to use RCTs as regulatory tools, it is necessary to provide some warrant of their external validity and impartiality. If we could have perfect causal knowledge of the effects of an intervention, impartiality would be warranted by default. But if there is uncertainty about it, RCTs should incorporate some warrants of impartiality.

[1] For a further discussion of the possibility of dispensing with randomisation in field experiments, see Deaton (2010) and Imbens (2010).

A regulatory decision should be impartial, and if we are going to ground it on inconclusive evidence, we need to make sure that nobody exploits such uncertainty in their own interest.

We have seen that the use of randomisation requires expert judgement, so the *mechanical objectivity* of RCTs is mere appearance: we need a subjective (judgement-based) assessment of its actual implementation in order to decide about the external validity of the trial. Nonetheless, randomisation provides a warrant of the impartiality of a clinical trial at the crucial stage of allocating treatments – not beyond that. Such warrant contributes to the credibility of the experimental outcome: we may question its external validity, but at least we can presume it is unbiased – at least more than unconstrained expert judgement.

As regulatory tools, RCTs have proven to be most successful at the FDA, where they are part of a system in four phases: the first two provide causal background knowledge for the trial, and the last one – post-marketing surveillance – controls for possible lacks of external validity. The number of label revisions, on the one hand, and market withdrawals, on the other, signals the levels of uncertainty with which the FDA is dealing. Randomisation, together with other means (such as blinding), has contributed to the impartiality of such uncertain regulatory decisions, making them more acceptable to the American public. However, it seems as though belief in the regulatory system has weakened today, and a debate has been started on how to strengthen the fairness of the FDA regulatory process.

If we are going to adopt RFEs as a public policy tool, we will probably need to work on two fronts. In Sect. 5, we have argued that randomisation needs to be successfully hidden from experimental subjects in order to be a warrant of impartiality in field trials, since we are not dealing with the biases of the researchers alone, but also with the preferences of the experimental subjects. On the other hand, as Cartwright has argued, and the example of the FDA seems to illustrate, a randomised trial per se does not warrant the external validity of its conclusions. We need to keep a record of the fallibility of the conclusions of field trials in order to measure the degree of uncertainty we are dealing with.

If we follow the institutional paradigm of the FDA, the question is how to integrate RFEs into an institutional system that makes their results credible. As of today, there is no clear answer as to which sort of institution should this be (Duflo and Kremer 2005). Government-sponsored programmes are rare because it is difficult to attain the high level of political consensus required for a successful implementation. Without this consensus, RFEs can easily fall prey to the sort of manipulations described in the previous section, in which each party will try to make the experiment support its views. Non-governmental organisations (NGOs) are more active, because they are interested in finding the most efficient way of spending their (usually scarce) resources and they are comparatively free to choose where and how they distribute them. However, NGOs create their own biases: the culture of the organisation implementing the assessment (e.g. the motivation of its employees) may impact on the participants' reaction in a way difficult to replicate in further extensions of the programme.

NGOs (or non-profit organisations in general for that matter) have also a problem of credibility, not unlike the pharmaceutical industry: they usually have a stake in the programmes they evaluate (Pritchett 2002). And randomisation does not seem to be a good enough warrant of impartiality to convince governments that they can trust an assessment and implement it at a bigger scale. This is probably why Duflo and Kremer (2005, pp. 115–117) advocate the creation of a sort of international 'regulatory agency' for development policies. International organisations involved in development should establish an office with the following mission. It should assess the 'ability of the evaluation to deliver reliable causal estimates of the project's impact' and 'conduct credible evaluations in key areas' (p. 115).

In other words, international organisations should provide the impartial expertise required to make the trials credible to the involved parties. This is probably the best solution. However, it remains an open question why would the participants in the trial see the international organisation as a neutral third party they can trust. Only if they do, one can be certain that the trials it sponsors are a credible source of knowledge about their target population.

Acknowledgements Our most sincere thanks to Hsiang-Ke Chao and Szu-Ting Chen for organising the very hospitable and intellectually fruitful conference in which this chapter was originally presented. Thanks to the editors and reviewers for their comments. Teira's research has been funded by the Spanish Ministry grant FFI2011-28835.

References

Akerlof, George. 1970. The market for 'Lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3): 488–500.

Banerjee, Abhijit V., and Esther Duflo. 2009. The experimental approach to development economics. *Annual Review of Economics* 1(1): 151–178.

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels. 2005. Fair procedures: Evidence from games involving lotteries. *The Economic Journal* 115(506): 1054–1076.

Carnevale, Anthony Patrick, Stephen J. Rose, and Century Foundation. 2003. Socioeconomic status, race/ethnicity, and selective admissions. http://www.tcf.org/Publications/White%5FPapers/carnevale%5Frose.pdf. Accessed 15 Jan 2012.

Carpenter, Daniel P. 2010. *Reputation and power: Organizational image and pharmaceutical regulation at the FDA*. Princeton: Princeton University Press.

Carpenter, Daniel, and Colin Moore. 2007. Robust action and the strategic use of ambiguity in a bureaucratic cohort: FDA scientists and the investigational new drug regulations of 1963. In *Formative acts*, ed. Stephen Skowronek and Matthew Glassman, 340–362. Philadelphia: University of Pennsylvania Press.

Cartwright, Nancy. 2007. Are RCTs the gold standard? *BioSocieties* 2(1): 11–20.

Cartwright, Nancy, and Eileen Munro. 2010. The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice* 16(2): 260–266.

Deaton, Angus. 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature* 48: 424–455.

Duflo, Esther. 2010. *La politique de l'autonomie*, Lutter contre la pauvreté, vol. 2. Paris: Seuil.

Duflo, Esther, and Michael Kremer. 2005. Use of randomization in the evaluation of development effectiveness. In *Evaluating development effectiveness*, World Bank Series on Evaluation and Development, vol. 7, ed. George Keith George, Osvaldo N. Feinstein, and Gregory K. Ingram, 205–232. New Brunswick/London: Transaction.

Duflo, Esther, and Rema Hanna. 2006. Monitoring works: Getting teachers to come to school: C.E.P.R. discussion papers, CEPR discussion papers: 5426. London

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. Using randomization in development economics research: A toolkit: C.E.P.R. discussion papers, CEPR discussion papers: 6059. London

Epstein, Steven. 1996. *Impure science. Aids and the politics of knowledge*. Berkeley: University of California Press.

Featherstone, Katie, and Jenny L. Donovan. 2002. "Why don't they just tell me straight, why allocate it?" The struggle to make sense of participating in a randomised controlled trial. *Social Science & Medicine* 55(5): 709–719.

Hacking, Ian. 1988. Telepathy: Origins of randomization in experimental design. *Isis* 79(3): 427–451.

Heckman, James. 1992. Randomization and social policy evaluation. In *Evaluating welfare and training programs*, ed. F. Manski and Garfinkel Irwin, 201–230. Cambridge/London: Harvard University Press.

Imbens, G. 2010. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48: 399–423.

Lasagna, L. 1959. Gripesmanship: A positive approach. *Journal of Chronic Diseases* 10: 459–468.

Macklin, Ruth. 2004. *Double standards in medical research in developing countries*, Cambridge Law, Medicine, and Ethics, vol. 2. Cambridge/New York: Cambridge University Press.

Marks, Harry M. 1997. *The progress of experiment. Science and therapeutic reform in the United States, 1900–1990*. New York: Cambridge University Press.

Marks, Harry M. 2000. Trust and mistrust in the marketplace: Statistics and clinical research, 1945–1960. *History of Science* 38: 343–355.

Orr, Larry L. 1999. *Social experiments: Evaluating public programs with experimental methods*. Thousand Oaks: Sage.

Petryna, Adriana. 2009. *When experiments travel: Clinical trials and the global search for human subjects*. Princeton: Princeton University Press.

Porter, Theodore M. 1995. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton: Princeton University Press.

Pritchett, Lant. 2002. It pays to be ignorant: A simple political economy of rigorous program evaluation. *Journal of Policy Reform* 5(4): 251–269.

Sackett, David, William Rosenberg, Muir Gray, Brian Haynes, and Scott Richardson. 1996. Evidence-based medicine: What it is and what it isn't. *British Medical Journal* 312: 71–72.

Stone, Peter. 2007. Why lotteries are just? *The Journal of Political Philosophy* 15(3): 276–295.

Teira, D. 2011a. Frequentist versus Bayesian clinical trials. In *Philosophy of medicine*, ed. Fred Gifford, 255–297. Amsterdam: Elsevier.

Teira, D. 2011b. *Impartiality in clinical trials*. London: University College London.

Urbach, Peter. 1985. Randomization and the design of experiments. *Philosophy of Science* 52(2): 256–273.

Wilson, Charles. 2008. Adverse selection. In *The New Palgrave dictionary of economics*, 2nd ed. Steven N. Durlauf and Lawrence E. Blume. London: Palgrave Macmillan.

Worrall, John. 2002. What evidence in evidence-based medicine? *Philosophy of Science* 69(3 Supplement): S316–S330.

Worrall, John. 2007. Why there's no cause to randomize. *The British Journal for the Philosophy of Science* 58(3): 451–488.

Chapter 12
Explaining the Explanations of 100 Million Missing Women

Hsiang-Ke Chao and Szu-Ting Chen

Abstract This chapter studies the methodology in the missing-women debate among economists and biologists. One of the central philosophical and methodological issues at stake in the missing-women debate is natural and social scientists' attempts for discovering the underlying causal structures and mechanisms. Although they encounter the same problem of inferring the mechanism and causal structure in face of available data, the discovering strategies vary. In this chapter, we will comparatively study the strategies of discovering causes and mechanisms in the case of missing women.

1 Amartya Sen's Missing Women

Nobel laureate economist Amartya Sen opened his 1990 *New York Review of Books* article, "More Than 100 Million Women Are Missing," with the following sentences:

> It is often said that women make up a majority of the world's population. They do not. This mistaken belief is based on generalizing from the contemporary situation in Europe and North America, where the ratio of women to men is typically around 1.05 or 1.06, or higher. In South Asia, West Asia, and China, the ratio of women to men can be as low as 0.94, or even lower, and it varies widely elsewhere in Asia, in Africa, and in Latin America. How can we understand and explain these differences, and react to them?

H.-K. Chao (✉)
Department of Economics, National Tsing Hua University,
101, Section 2, Kuang Fu Road, 30013 Hsinchu, Taiwan
e-mail: hkchao@mx.nthu.edu.tw

S.-T. Chen
Graduate Institute of Philosophy, National Tsing Hua University,
101, Section 2, Kuang Fu Road, 30013 Hsinchu, Taiwan
e-mail: stchen@mx.nthu.edu.tw

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_12,

Sen referred to the observations of exceptionally low female-male population ratios in certain Asian countries compared with those in Western countries. Recently, this issue has been referred to in the literature as the "missing-women" problem. It is a problem, as Sen sees it, because the observations reveal an abnormality. The normal circumstance would consist of the following elements. At the outset, the laws of nature suggest that the number of male births is higher than female births; therefore, the men-women sex ratio at birth should be higher than unity.[1] But the mortality of males is also higher than that of females, a fact that is thought to be compensated for by the higher male birth rate—a natural "regulatory mechanism" that should result in an actual men-women ratio of unity (Chahnazarian 1988, p. 217). This suggests what the observed sex ratio should be if biological factors were the only ones to affect human reproduction.

However, for Sen, the *natural laws* are conditional on gender equality; that is, they are conditional on whether men and women receive similar nutritional and medical attention and health care. Consequently, when the observed sex ratio deviates from that which biology would indicate in such a way as to favor the number of men, then there are missing women, and that calls for explanations and calculations. Moreover, such a deviation should be attributed to the unequal social and cultural treatment of women. Sen argues that these sociocultural causes could be expressed in the single composite factor of *son preference*, meaning the parental preference for boys over girls. Sen subsequently concludes that there are 100 million women missing due to the prevalence of son preference in the non-Western countries—50 million in China alone.

There are some methodological issues concerning Sen's assertion. On the one hand, there is the question of exactly how many women are missing; on the other hand, there is the question of the cause of this phenomenon. The answers to these questions are dependent on each other. To calculate the number of missing women, we must determine the numbers of males and females in an entire population to calculate the sex ratio, which in turn is affected by factors that cause either low birthrates, high mortality rates, or shortened life expectancy of females. Sen thus provides a possible explanation for the way the factors categorized as son preference result in missing women.

Sen's study reflects the long-standing issue of the human sex ratio, which has been studied since the seventeenth century, when statistical data were first collected. For example, John Graunt (1662) and John Arbuthnot (1710) statistically analyzed the English data and showed that the high male-female sex ratio at birth (around 105 boys per 100 girls) was not due to chance (Hacking 2006). Evolutionary theories of sex ratio were developed in the nineteenth century (Sober 2007). These studies provide the foundation for the *normal*, or *natural*, state, in which nonnatural factors do not intervene with the sex ratio at birth.

[1] Sex ratio at birth is also called *offspring sex ratio*. Throughout this chapter, the sex ratio is defined as the number of men per women.

Based on this work, social scientists conducted research on the socioeconomic mechanism behind and the causes of the missing women. In Chahnazarian's (1988) extensive review of the biological and sociodemographic determinants of the sex ratio at birth, the latter includes factors such as maternal age, paternal age, and birth order. In addition, economists have investigated the explanatory power of economic factors such as income and wage. They go even further to study whether and how son preference is affected by such factors. Sen, for instance, implied that the anomalous sex ratio may have been caused, and might be cured, by the economic factor "gainful employment," defined as the condition of having paid work outside the household. Women's gainful employment may have an effect on other factors—such as son preference and women's inequality—and in turn on the sex ratio.[2] Hence, Sen urges policy interventions, such as providing better education for women, as a cure.

In this chapter, we review the debate about the cause of the missing-women phenomenon—that is, the debate about whether the phenomenon is caused by biological or cultural determinants. Our purpose is to show a crucial role that the conception of causal structure plays in explaining or explaining away the supposed observation of missing women. It is also argued that the complementary strategy has been adopted to form a relatively more complete causal structure that can be used to tell a relatively more complete causal story about the underlying mechanism that is thought to be responsible for the occurrence of Asian countries' missing women.

2 How Baruch Blumberg Explains His "Strangest Observation"

In biomedical science, researchers attempted to seek different explanations for the missing women. One example is Baruch S. Blumberg, the 1976 Nobel laureate in medicine. Blumberg's research focuses on the hepatitis B virus (HBV), which he discovered in 1965. He observed during his fieldwork in Greece possible associations between HBV infection and sex ratio (Hesser et al. 1975). He later found similar results in the countries having high HBV prevalence, such as Greenland, Kar Kar Island, and the Philippines.[3] Specifically, Blumberg found that "carrier" families, that is, either parent was a chronic carrier who is a person who tests positive for hepatitis B surface antigen (HBsAg) but does not develop an antibody against HBsAg (anti-HBs), had a higher sex ratio (i.e., more boys) than did the families that had no evidence of HBV infection. In contrast, the "antibody"

[2] See Qian (2008) for an empirical study of the relation between women's gainful employment and sex ratios.

[3] Oster (2005, Table 3) summarizes the original data of these micro-studies of offspring sex ratio by parental HBV infection.

families (either parent had anti-HBs) had the lowest sex ratio. Moreover, Blumberg found that HBV carrier mothers have fewer female births and consequently conjectured that HBV may cause a high sex ratio at birth.

The observation of the relation between the response of a parent to infection with HBV and the gender of his or her offspring is Blumberg's "strangest" one, because it means that HBV is gender biased (Blumberg 2002, p. 182). In a sense, there is biological evidence supporting Blumberg's observation. Before Blumberg proposed his hypothesis, scientists had already known that HBV infections are gender sensitive. For instance, males are more likely than females to become HBV carriers, that is, more likely to be HBsAg(+), whereas more females than males develop anti-HBs. As a result, Blumberg thought that he had provided a "biological explanation" for the issue of the missing women, an issue he had learned about from a prominent demographist, Ansley J. Coale:

> [Coale] published a paper concerning the high sex ratios that have been observed in China. He proposed that, because there is no other biological explanation, the apparent deficit in female births in China could be a consequence of female infanticide. As might be expected, his findings and conjectures had a big play in the media. I pointed out to him that there might be a biological explanation. China, and particularly South China, has some of the highest frequencies of HBV carriers in the world. If our observations on the relation between carriers and gender of offspring in Greece and elsewhere were also valid in China, then this might provide a biological explanation for apparent loss of female children. (Blumberg 2002, p. 185)

Carrier families are more likely than antibody families to have male children. Consider the countries where the incidence of both HBV and son preference is high. If families would like to have children until they reach the desired number of boys, then antibody families, who are more likely to have girls, would have a lower sex ratio and a larger family size than carrier families. The hepatitis B hypothesis hints at one intriguing policy implication: if an HBV vaccination program were successful, then there would be fewer carriers and hence a lower sex ratio at birth, with the consequent effect on family size and the sex ratio of the population.

Drew et al. (1978) first offered a theory on the way HBV affects sex ratio. Based on the facts that males are more likely to be HBsAg(+) and females are more likely to be anti-HBs(+) and the statistical facts that parents with HBsAg(+) have a higher sex ratio (more boys) and parents with anti-HBs(+) have lower sex ratio (more girls), Drew et al. (1978, p. 691) conjecture that males would recognize HBsAg as "self" and remain HBsAg(+) persistently (i.e., become carriers). By contrast, females would be more likely to recognize HBsAg as "foreign" and thus produce anti-HBs. Hence, HBsAg(+) is associated with the possibility of giving birth to more boys, and anti-HB(+) is associated with giving birth to more girls. Drew et al. (p. 691) provided some possible explanations: either HBsAg protects male fetuses or anti-HBs hinders fertilization by a Y-bearing sperm. But the proposed mechanism is still a conjecture, as Blumberg recently admitted that there is no evidence for such effects of HBV in utero (2007, p. 229).

3 How Economists Explain 100 Million Missing Women

Blumberg's hepatitis B hypothesis is interesting, but no economist took it seriously until Emily Oster's (2005) article was published. Oster tested Blumberg's idea on a population level using both time-series and cross-sectional data. In particular, Oster used the historical data of Taiwan's universal vaccination of all newborns beginning in 1984 to conduct an empirical test to see whether there is indeed a significant positive correlation between the variable of HBV prevalence and an increase in the sex ratio at birth. Taiwan's vaccination case forms a natural experiment that is ideal for testing the correlation, because economists can analyze the historical data of births to vaccinated and unvaccinated mothers to see whether there is a great gap in offspring sex ratio between these two groups; if there is indeed a significant difference in sex ratio at birth, then HBV has a positive effect on the percentage of children who are male; otherwise, it does not.

Although Oster's testing result showed that there is a significant correlation between HBV prevalence and sex ratio at birth, she noted a caveat to her conclusion. Oster observed that Taiwan's vaccination program coincided quite closely with an increased availability of fetal sex-determination technology and a probable increase in sex-selective abortion. Consequently, the magnitude of the difference in sex ratio at birth between vaccinated and unvaccinated mothers is likely to be smaller than it otherwise would be, because the effect of vaccination on decreasing sex ratio at birth is offset by the countervailing behavior of adopting fetal sex-determination technology, which was motivated by the cultural cause of son preference.

However, in a cross-country analysis, Oster, by applying the least-square method on single equations, found a significant correlation between HBV and sex ratio at birth. She also stated that about 75 % of Sen's 100 million missing women could be explained by parental infection with HBV, implying that son preference plays a lesser role in explaining the missing women. Soon after, Blumberg (2006a, b) characterized Oster's finding as one of the great achievements of HBV research, because it supported his previous findings and confirmed that HBV does have effects on gender.

In response to Oster's biological explanation, Monica Das Gupta, a supporter of the cultural explanation, maintained that the Chinese sex ratio at birth for the first birth was always within the normal range of 1.05–1.06. A higher sex ratio at birth was observed only in subsequent births. In addition to these two empirical testing results, Das Gupta found that an extremely high sex ratio at birth was observed mainly among women who had previously given birth only to daughters. Together, these results strongly suggest that it is son preference, rather than HBV, that has the significant effect on distorting sex ratio at birth and therefore it is the former, rather than the latter, that is the cause of the missing women (Das Gupta 2005, 2006). In response, Oster (2006) pointed out that she does not disagree with the cultural explanation; rather, she opposes Das Gupta's position that "the support for cultural explanations allows to conclude that the biological explanation is not particularly

salient" (Oster 2006, p. 324). To reconcile her biological explanation with the cultural one, Oster, with her coauthor, proposed a further hypothesis: it is possible that HBV interacts in a complicated way with birth order and the sex of previous children (Oster et al. 2008).

Against this background, a pair of Taiwanese economists—Ming-Jen Lin and Ming-Ching Luoh—conducted a study that was designed mainly to test whether Oster's further hypothesis is sustainable (Lin and Luoh 2008). They acquired a unique dataset of Taiwan's nationwide hepatitis B vaccination program that was launched in July 1984. The data, more substantial than Oster's Taiwanese data in size, are the comprehensive historical data obtained from the Hepatitis B Mass Immunization national databank of Taiwan, which includes gender, year and month of birth, birthplace, mother's age, birth order of the child, unique ID of the mother, and mother's HBV status at the time of pregnancy. Lin and Luoh conducted a regression analysis to see whether there was any significant correlation among gender, birth order of the child, and mother's HBV status. The specification of their regression is as follows:

$$\begin{aligned}\text{Boy} = \alpha &+ \beta_1\ \text{HBsAg} \\ &+ \beta_2\ (\text{Birth Order 2}) \\ &+ \beta_3\ (\text{Birth Order 3 and higher}) \\ &+ \beta_4\ \text{HBsAg} \times (\text{Birth Order 2}) \\ &+ \beta_5\ \text{HBsAg} \times (\text{Birth Order 3 and higher}) \\ &+ \beta_6\ (\text{Mother Age Dummies}) \\ &+ \beta_7\ (\text{Child Birth Year Dummies}) \\ &+ \beta_8\ (\text{Birth Township Dummies}) \\ &+ \varepsilon \end{aligned} \tag{12.1}$$

Parameter β_1 measures the effect of HBV on sex ratio at birth; parameters β_4 and β_5 investigate whether the effect of HBV differs among different birth orders. Lin and Luoh's argument appeals to both substantive and statistical significance. If the testing results are that the estimated value of β_1 is small and statistically insignificant and if there is no significant difference between β_4 and β_5, it then suggests not only that HBV plays no role in determining sex ratio at birth but also that some part of the further hypothesis—that is, the hypothesis that there are complex biological mechanisms in operation between HBV and birth order—should be ruled out. On the other hand, the magnitude of the values of β_2 and β_3 indicates whether the offspring in higher birth orders are more likely to be male; if the magnitudes are large enough to be significant, then the result supports the son-preference explanation. The end result of Lin and Luoh's testing is that β_1 is small and sometimes insignificant; measuring the effect of HBV on sex ratio at birth, parameters β_4 and β_5 are small and insignificant, confirming that birth order does not amplify the effect of HBV. Together, these two results show that the biological explanation is

not sustainable. As for β_2 and β_3, the value of β_2 is small and insignificant, and β_3 is very large and significant; these results together indicate that variation in the sex ratio at birth mainly affects third and later children and thus suggests that the cultural explanation is more plausible.

To test the other part of the further hypothesis—that the effect of HBV varies with the sex of previous children—Lin and Luoh ran another regression as follows:

$$\begin{aligned}\text{Boy} = \alpha &+ \beta_1 \text{ HBsAg} \\ &+ \beta_2 \text{ (First two children are girls)} \\ &+ \beta_3 \text{ HBsAg} \times \text{(First two children are girls)} \\ &+ \beta_4 \text{ (Mother Age Dummies)} \\ &+ \beta_5 \text{ (Child Birth Year Dummies)} \\ &+ \beta_6 \text{ (Birth Township Dummies)} \\ &+ \varepsilon \end{aligned} \tag{12.2}$$

Again, following the same logic of identifying whether the effect of HBV varies with birth order, Lin and Luoh checked whether the values of β_1 and β_4 were large and significant; if they were, the result would indicate both that HBV indeed plays an important role in determining the offspring sex ratio and that there is indeed some interaction between HBV and the sex of the first two children. According to Lin and Luoh's result, both β_1 and β_3 are small and insignificant, but β_2 is very large and significant. This result thus also suggests that the son-preference explanation is more plausible. So this allows the authors to rule out the "complex biological mechanisms" of HBV inflections on the sex ratio (Lin and Luoh 2008, p. 2264).

Oster soon struck back. She teamed up with Blumberg (Blumberg and Oster 2007). They argue that, empirically, paternal, not maternal, hepatitis carrier status is more strongly correlated with sex ratio at birth. Yet their 13-page manuscript was never completed and published. Perhaps Oster found that the hypothesis was not sustained; her recent 2010 paper, written with three Chinese medical officials and researchers (Oster et al. 2010), claims that by analyzing the data of Haimen City in Jiangsu Province, China (sample size = 67,511 individuals), no relationship between paternal HBV and the missing women is found. Consequently, she conceded that HBV cannot explain the missing women in Asia.

4 Difference in Methodology

The case of missing-women debate can be seen as an example of *extrapolation* in the sense of Daniel Steel (2008; see also Steel's chapter in this volume, 2013). Researchers compare the target with the sources originating from other disciplines and geographic regions and draw similarities that could serve as a guide for further investigation. Where the strategy of discovering the missing women is concerned,

we notice methodological similarities and dissimilarities among economists and biomedical scientists. First, we find that Sen's and Blumberg's theorization processes may best be understood in terms of *Inference to the Best Explanation* formulated by, inter alia, Peter Lipton. Lipton's Inference to the Best Explanation has a root of Charles Sanders Peirce's *abduction*, which consists the form of inference: "The surprising fact, C, is observed; but if A were true, C would be a matter of course, hence, there is reason to suspect that A is true." Inference to the Best Explanation stresses also on inductive inference. Lipton maintains that scientists infer from available evidence to the hypothesis that would provide the best explanation for that evidence. For instance, as Lipton states, Darwin inferred from his biological evidence the theory of natural selection because natural selection would best explain that evidence (Lipton 2000, p. 184). Conversely, inference is guided by explanatory considerations, particularly by those explanations that provide most understanding if true—the *loveliest* explanations.[4] This seemingly circular argument in fact describes well how a hypothesis is formulated in science, how it is justified as the best hypothesis among many competing hypotheses, and how it can heuristically guide the inference process. It is also stressed that Inference to the Best Explanation is fallible. Because all available evidence does not necessarily lead to truth, the best explanation may not be the actual explanation. Lipton's reconstruction of Carl Hempel's (1966) discussion of Ignaz Semmelweis's attempt to explain different rates of childbed fever in two hospital wards well demonstrates how the best explanation to such observation can be derived from the account of Inference to the Best Explanation (Lipton 2005).

It seems both of Sen's and Blumberg's studies fit well the account of Inference to the Best Explanation. In our case study, Sen and Blumberg seek to infer from evidence the best explanations for Asia's missing women. Sen derives missing-women phenomenon from the sex ratios in Asia from those in the Western countries, and then he infers from the evidence that the son preference induced by women's gainful employment is the best explanation for Asia's missing women. Similarly, Blumberg infers from his data that the hepatitis B hypothesis is the best biological explanation for the missing women, in which he uses his knowledge of the hepatitis virus, especially the feature of gender sensitivity. Despite the fact that both Sen's cultural hypothesis and Blumberg's hepatitis B hypothesis provide the best explanation for their evidence, we observe that the consideration of explaining the missing-women phenomenon has been the guiding force to direct them to adopt a particular inference strategy for carefully developing their accounts by explaining and explaining away the adequacy of explanatory components. So doing makes their explanation better than others. Yet it would still be understandable should Sen's and Blumberg's hypotheses have turned out to be false explanations in the face of newly acquired data, and have been replaced by a theory providing better account for the data.

[4] By contrast, a likeliest explanation is the explanation that is best warranted by the evidence.

Contrasting with Sen's and Blumberg's attempts to theorize the missing-women phenomenon, Oster's and Lin and Luoh's studies are conducted by running regressions on data to find significant statistical relations between high sex ratio at birth and factors that might influence it. Although they intend to find the causal power of the factors, it is not obvious whether they are concerned with finding the true causal relationships. For instance, Oster (2005, p. 1164) stated that "after one adjusts for differences in the sex ratio at birth *caused* by hepatitis B, the number of missing women (based on population estimates from 1980 to 1990) drops to 32 million" (our emphasis), but she has been cautious in using causal language in other places. In Blumberg and Oster (2007), she used correlation ("paternal, not maternal, infection is correlated with higher offspring sex ratios") more often than causation. Oster is concerned only to justify the hepatitis B hypothesis by claiming that the causal relation is supported by relevant correlations. Similarly, although Lin and Luoh found that cultural factors are causal to the missing women and they claimed that the biological mechanism is ruled out, they could not, or would not, find the true causal mechanism.

We might use Trygve Haavelmo's (1944) famous mechanical analogy to illustrate. We could derive the functional relation between the pressure on the throttle and the speed of a car on a flat road under usual conditions, but such a relation will break down as soon as there is a change in any working part of this car, or a change in an external condition. The throttle-speed relation is less *autonomous* because this type of relation is not invariant to changes in the surrounding conditions and thus is not fundamental to economics. For Haavelmo, the general laws of thermodynamics and the dynamics of friction are examples of highly autonomy relations, because they "describe the functioning of some parts of the mechanism irrespective of what happens to some other parts" (Haavelmo 1944, p. 28). The real automobile mechanism is hard to be discovered without opening the hood, one might still be able to find out what causes the throttle/speed relationship to break down. The econometric studies of Oster and of Lin and Luoh focus on economic relations such as the one between throttle pressure and speed; they are interested in whether or not certain relations sustain, rather than opening the hood and seeing the engine of the car. They intend to find factors that causally relate to sex ratio at birth, empirically speaking, and use the empirical finding to vindicate or repudiate hypotheses.

5 Causal Structure and Net Result

Given Lin and Luoh's empirical result that the effect of hepatitis B on an increase in the sex ratio at birth is limited, we might conclude that economists are no longer willing to endorse the idea that hepatitis B has the efficiency to affect the sex ratio.

This description seems to be further justified by Oster's remarks about the idea of science:

> I'd be lying if I told you it wouldn't be great if I was right all the time. . .If you work like this, especially if it's something that people care about, and you get to collect some more data that is maybe going to be even more informative than what you had before, it's your responsibility to do that. This is the way science works.[5]

On the face of it, Oster's statement might be regarded as representing two methodological points. One is the doctrine that follows the tradition of Popperian falsificationism; the other is that the size of data does matter. The aim of science, including economics, is to identify stable connections—or *regularity laws*—among variables (or factors) of interest so that, by using these connections or laws, scientists are able to explain the phenomena that are thought to be governed by the connections or laws. If, in the process of hypothesis testing, the scientists somehow find out that there are no stable connections or lawlike relations among the targeted variables, the correct action is for scientists to repudiate their hypotheses about those connections or laws. In our case, the hepatitis B hypothesis is that there is a positive connection between the prevalence of hepatitis B among mothers of newborns and the increased ratio of newborn males to females and thus the number of missing women. In her 2005 paper, Oster, by using various quantitative strategies and a larger dataset than Blumberg's, concluded that there is indeed a robust connection between hepatitis B and the missing women, so she suggested that hepatitis B could explain the phenomenon. Later, Lin and Luoh's study of three million newborns rejected the hepatitis B hypothesis. In order to respond to Lin and Luoh, Oster et al. (2010) used new empirical data and found no effect of hepatitis B carrier status on the sex ratio. This finding led her to reject the hepatitis B hypothesis and claim that hepatitis B does not explain male-biased sex ratios in China. Oster's practice of rejecting her previous hypothesis thus illustrates that her action following her doctrine of science.

However, if we interpret the missing-women debate in this methodological sense, then the following short paragraph, which is quoted from the 2008 working paper version of Oster et al. (2010), may seem puzzling (Oster et al. 2008, p. 6):

> [After showing] that hepatitis B carrier rates cannot explain male-biased sex ratios or the 'missing women' in China . . . [a]n important remaining issue is whether it is possible to reconcile the biological results in the original paper (Oster 2005) with these [2008] results and, in particular, how the individual-level data from outside of China and the evidence from vaccination campaigns in Alaska can coexist with new results from China . . . We revisit the original individual-level data from Greece and the Philippines and continue to find support for the connection between *paternal* hepatitis B carrier status and offspring sex ratio. Moreover, in the data from China discussed here, we also find some interaction between hepatitis B, gender and fertility: women with the hepatitis B *e* antigen (carriers who are also replicating an additional viral antigen) seem to have fewer male children. Further, women who are carriers of the virus have fewer children overall, even with extensive controls. Together, this evidence suggests that there may still be some interaction between hepatitis B and fertility outcomes (in general) but that clearly the pathways are much more complicated than the simple carrier-male offspring connection.

[5] Quoted in Justin Lahart, "Economist Scraps Hepatitis Theory on China's 'Missing Women,'" *The Wall Street Journal*, May 22, 2008.

If we read between the lines, the quotation seems to suggest that even though the new empirical result does not support the hepatitis B hypothesis, the hepatitis hypothesis should not be rejected immediately; economists would rather at one stage hang on to the refuted hypothesis and try to find a way to reconcile the old findings with the new ones. In this sense, we might define two methodological views of science. On the one hand, science is traditionally regarded as being constituted by a great number of theories. The aim of each theory is to provide an explanation for a phenomenon in question. Each theory in turn contains a number of hypotheses, each of which posits a general law that is supposed to govern or regulate a corresponding part of the targeted phenomenon if that part can be logically derived from the general law. Once all the relevant hypotheses—that is, the relevant regularity laws—of the theory can be used to derive the corresponding parts of the phenomenon, the theory is said to provide an explanation of the phenomenon. When a new phenomenon of a similar kind can no longer be derived from—or explained by—the same theory (the same set of regularity laws), then the theory and its component laws should be *substituted* or *replaced* by other new theories and their component new laws. We may call the idea that involves the description of scientists' practices the *substitutive* conception of scientific practices. This is Popperian in spirit.

On the other hand, our case study shows that the substitutive framework does not necessarily appear in economists' practices. The empirical result of Lin and Luoh is thought to reject the hepatitis B hypothesis by their findings of little effect of HBV on the sex ratio at birth and a significant correlation between higher birth order and the sex ratio at birth. Yet we also witness the so-called *complementary* conception of scientific practice as followed in Oster's work: when scientists want to check whether there is a significant correlation between any two targeted events, they usually apply empirical tools, such as regression analysis, to run a significance test. When they find that the estimate that represents the correlation does indeed fall within the prescribed confidence interval, they accept the correlation hypothesis and confirm that there is a correlation between the two events. In contrast, if the estimate falls outside the confidence interval, it is normally presumed that such a hypothesis should be rejected and no correlation is acknowledged. Yet in practice it is intended to append additional conditions to the failed hypothesis to explain why the hypothesis conflicts with the testing result. In other words, some explanation is provided to reconcile what is stated in the hypothesis and the contradictory testing result. In fact, Oster herself used the term complementary that fits precisely into our observation. In her reply to Das Gupta's comments, Oster stated (Oster 2006, pp. 325–326),

> The key to thinking about the relative potential of culture and biology to explain the over-representation of men in a population is understanding that marginal effects may be seen to operate and still tell us relatively little about the average. In the end, it seems better to think of these two explanations as *complementary*. The issue of gender imbalance in Asia—the causes and consequences—is an important one; we should endeavor to have a *complete* understanding, not just a partial one. (our emphasis)

In order to have a complete understanding, it needs first to understand the fact that the observed correlation between any two targeted events is in fact the *net result* of a complicated interaction among a great many relevant factors involved in

the two events. Based on this supposition, thinking in terms of causal structure is helpful. Hypothesis testing is conducted under an overarching assumption that the relevant factors interact with each other to form a *causal structure*, defined by Nancy Cartwright as "a fixed (enough) arrangement of components with stable (enough) capacities that can give rise to the kind of regular behavior that we describe in our scientific laws." Cartwright (1997, p. 343) explains that "capacity is used to mark out abstract facts about economic factors: what they would produce if unimpeded" (Cartwright 1998, p. 45). According to the idea of causal structure, it may seem that what is supposed to be stable are the capacities (or causal powers) possessed by those member factors of the structure. However, the regular association of the targeted factors that is described in the regularity law is generated from a causal structure, and therefore the stability of the association—that is, the degree to which the law is regular—depends on whether the causal structure itself is in a stable condition. If, however, an unexpected new factor intrudes into the original layout of the structure, this new factor will destroy the original stable condition of the causal structure and therefore bring about a new equilibrium state that is not consistent with what is stated in the original regularity law. From this perspective, Cartwright argues that, contrary to the general belief that regularity laws are necessary regular associations that regulate or govern any two targeted factors, "[regularity] laws are transitory [or contingent] and epiphenomenal, not eternal"; the stability of regularity laws is subject to whether the causal structure is capable of repeatedly generating them (Cartwright 1999, p. 122).

Based on this structural thinking, it may seem that a regularity law should be repudiated when the result of hypothesis testing tells us to do so. But we shouldn't throw out the baby with the bath water; we shouldn't throw out the causal factors with the defective regularity laws, but rather should attribute the failure of their hypothesis to the instability of the causal structure rather than to the causal factors, which possess stable capacities or causal powers. Whenever there is an anomalous phenomenon that is inconsistent with their hypothesis about a regular correlation between two factors, we can still consult the old knowledge of the capacities of the relevant factors and then *imagine* a new causal structure from which, by using these two pieces of information, a new regularity law can be derived. The new regularity law can in turn be used to explain why and how the previously unexplainable phenomenon occurs. The old knowledge of capacities is in this sense complementary to the knowledge of the new causal structure and regularity law. We may then call this description of economists' practices the complementary conception of scientific practices.

6 Strategies for Discovering Asia's Missing Women: Causal Structures

Let's again use the case of missing women to illustrate how economists develop complementary scientific practices. The hepatitis B hypothesis states that HBV has a positive effect on sex ratio at birth (SRB). This simple causal structure is illustrated in Fig. 12.1.

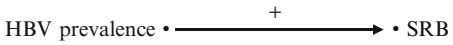


Fig. 12.1 The simple causal structure of hepatitis B hypothesis

where the symbol "+" indicates positive effect.

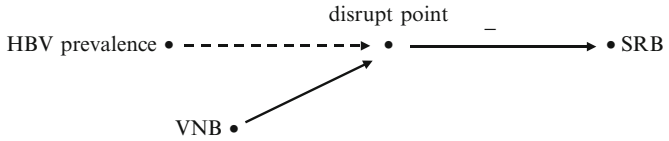


Fig. 12.2 The biological causal path being disrupted by vaccination campaign

where the symbol "−" denotes negative effect; the dotted line denotes the disruption of the original effect.

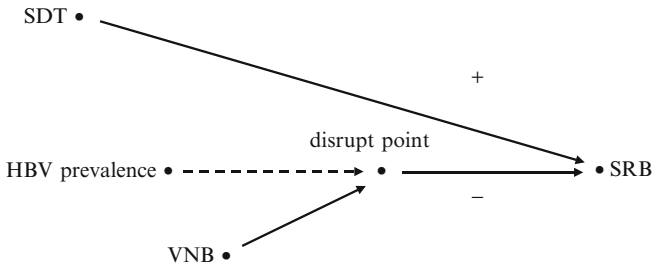


Fig. 12.3 The relatively more complicated causal structure with the further causal influence of SDT

Figure 12.2 illustrates that, because of the introduction of a universal vaccination program for newborns (VNB), the original causal path is disrupted by the intrusion of VNB, and the figure also shows that VNB has the negative effect of reducing SRB.

Figure 12.3 shows that, by further considering that Taiwan's universal vaccination program was launched about the same time as the use of fetal sex-determination technology (SDT) became prevalent, the original causal structure should be supplemented with an additional causal path leading from SDT to SRB; the figure shows that SDT has the positive effect of increasing SRB.

In Fig. 12.3, there are two competing causal paths that jointly determine the degree of SRB, and the entire causal structure represented in Fig. 12.3 also reflects the concern that, because of the introduction of SDT, the magnitude of VNB's effect on reducing SRB is likely to be offset by the new effect of SDT, which has a positive effect of increasing SRB to a certain degree.

It is mentioned in Oster (2005, p. 1185) that this offsetting effect would be mitigated if a large portion of the population were indifferent to whether they have sons or daughters. With the addition of this further concern, the causal structure

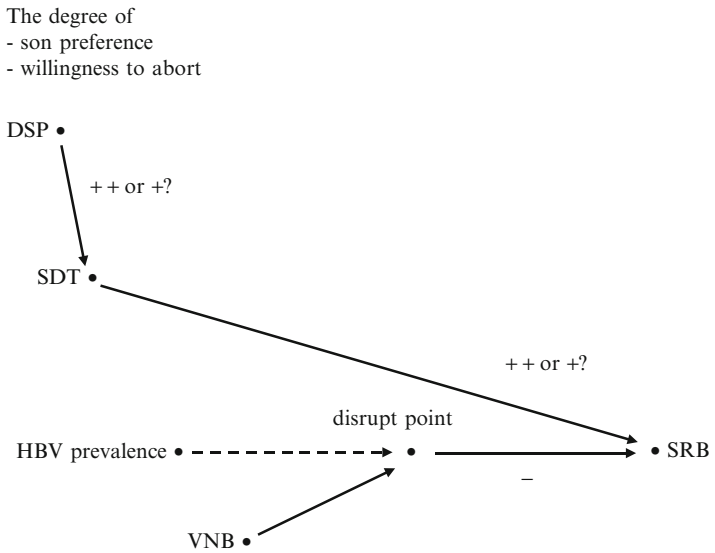


Fig. 12.4 The relatively most complete causal structure with the cultural and biological causal paths

depicted in Fig. 12.3 is revised into the new structure represented in Fig. 12.4. Depending on whether the population has a high or a low degree of son preference, SDT will have very strong positive (++) or mild positive (+) effect on increasing SRB. In other words, if a high percentage of the population prefers having sons, the VNB's effect of decreasing SRB would be much lower than it would otherwise be. In contrast, if most people in the population are indifferent to their offspring's gender, then the indicators estimated from Taiwan's vaccination campaign representing the correlation between HBV prevalence and SRB likely would more reliably reflect the VNB's undisturbed magnitude of the effect on decreasing SRB.

Note that the upper part of Fig. 12.4 represents the cultural and social causal paths that are formed by the connections of the cultural causes and would have a certain degree of influence on SRB, and the lower part of the figure consists of the biological causal paths that are shaped by the combinations of the biological causes and would also have a certain degree of effect on SRB. Also note that, in the process of building the relatively most complicated causal structure represented in Fig. 12.4, at each step an additional causal path is appended to the original relatively simple causal structure to form a relatively more complex figure; one important characteristic of the theory-building methodology is that, in the end, both the cultural and the biological groups of causal paths seem to be accommodated by the relatively most complicated causal structure to form the relatively most complete causal story. It is obvious that, in our description of the methodology, no win-lose strategy is applied in choosing between two competing causal stories. What it shows is that these two competing causal stories complement each other to form a more complete causal story.

This complementary methodology can also be illustrated by the debate between Das Gupta and Oster. According to Das Gupta, the change of sex ratio at birth is so closely correlated with the sex composition of the existing children in the family that it is very unlikely that biological factors could play any significant role in determining that ratio; she therefore concluded that the cultural explanation should replace the biological one. In response, Oster provided a miniature formal model to show that, regardless of the naturally occurring average, resource-constrained utility-maximizing parents would still behave in a way that has the marginal effect of changing the sex ratios (Oster 2006, pp. 326–327). In addition to the formal presentation of the difference between the average and marginal effects, Oster gave an example to illustrate her main point. Consider that there are two countries: country A is in the desert, and country B is in the Arctic. On average, due to its location, country A is hotter than country B. Imagine that we also observe that country A is cooler when the weather is cloudy, and country B is hotter when the weather is sunny. It is obvious that we won't therefore conclude that the entire difference between the weather of country A and that of country B is cloud cover; instead, we would say that "there is a *naturally occurring difference* in the *average* temperature but on the *margin* the temperature in both places can move" (Oster 2006, p. 325; emphasis added).

Thus, we can interpret Oster's example in terms of Cartwright's notions of causal structure and capacity. HBV can be regarded as a factor possessing the capacity to produce a naturally occurring difference—that is, a change in sex ratio at birth—if unimpeded. In the actual world, economists surely know that no factor can operate in an undisturbed environment. They, however, can *assume* that the various effects of these disturbances can be averaged out so that the phenomenon of interest—normally a relation between two targeted factors—can be observed *as if* the observed phenomenon were undisturbed. By using the ideas of capacity and causal structure, we can thus describe the positive relation between HBV (targeted factor 1) and sex ratio at birth (targeted factor 2) as a net result produced by HBV's capacity under a stable causal structure; the various effects of the relevant factors are all averaged out except the two targeted factors. According to such thinking, we can thus suppose that HBV, on average, can have a positive effect on the sex ratio.

Next, by following the logic illustrated in Oster's weather analogy, let's suppose that there are two countries: country A, which has a high percentage of HBV-infected population, and country B, which has low rate of HBV infection. And let's further suppose that we observe that people in country A have no tendency for son preference, but people in country B, for local reasons (such as the need to have more males to do hard agricultural work), prefer having more male births. We can then expect that when the families in country B already have two girls, they will try their best—for example, they may try fetal sex-determination technology—to have a boy in their third birth. As a result, in country B, we can observe a high correlation between the sex of the previous children in a family and the sex ratio of subsequent births. Therefore, just as we may perceive in the weather example that "there is a naturally occurring difference in the average temperature but on the margin the temperature in both places can move," we can also observe in the HBV case that

there is a naturally occurring difference in the average sex ratio (country A, due to the effect of HBV, on average has a higher sex ratio than country B) but on the margin the sex ratio in both places can move (country B, due to its own local concern, may have a higher sex ratio than it would normally have). In the case of country B, we observe that there is a negative relation between HBV and sex ratio, a result that is contradictory to what is described in the capacity claim of HBV (HBV is a factor possessing the capacity to produce a positive change in sex ratio). However, just as we have noted in the discussion of causal structure and capacity, this paradoxical conclusion can be reconciled by resorting to a change in the original structure. Our original simple causal structure contains only one causal path: the HBV infection rate of country B leads directly to the sex ratio at birth of country B. To accommodate the new fact that country B is a son-preference country, we need to add a new causal path leading from country B's degree of son preference to its sex ratio at birth. These two causal paths together constitute a more complicated new causal model from which a new net result—a negative association between HBV and the sex ratio—is generated.

7 The Causal Images of Mechanisms

We have identified two methodological approaches in the missing-women debate: substitutive and complementary. We have also argued that the complementary approach can be understood as an attempt to identify the causal structure. Studies that seem to significantly conflict (such as Blumberg and Oster 2007) with the test that is regarded as decisive (i.e., Lin and Luoh 2008) may be interpreted as attempts to find the strengths of causal paths. Because it is the net result of the causal structure that the empirical test tests against, the data cannot reject the existence of the causal structure. All in all, there is an image of a causal structure in scientists' minds.

Because of the influence of the received image of science, we often find that empirical economists refer to correlation and regularity but seldom discuss mechanism. In the sense of Machamer et al. (2000), we can say that the genuine underlying structure that produces the phenomenon of the correlation between HBV and sex ratio at birth—be it a positive or a negative relation—is the mechanism that makes the relation what it is. When the manifest regularity between HBV and sex ratio at birth is a positive relation, it is the mechanism that makes it so; when the regularity is a negative one, it is still the mechanism that makes it so, but, this time, the mechanism operates under a different structure that makes the originally positive relation become negative. By ignoring the underlying mechanism, the biological explanation is ruled out simply because a robust relation between HBV and sex ratio at birth can no longer be found. However, a faint image of causal structure—like Machamer, Darden, and Craver's mechanism schemata and sketches of the underlying mechanisms—is still possessed by anyone who is interested in the missing-women phenomenon. We can infer a great many possible

situations that may correspond to the causal structure of the sea of disturbances; then, based on this partial causal image, we thus append a complementary causal story to our original biological explanation, as in our case of the missing women, to make the explanation more complete and plausible.[6]

This patch-up methodology is not conducted for the purpose of preserving a falsified hypothesis; instead, the methodology is supplemented with the process of theory-building in an attempt to obtain—as illustrated in our case—a more complete understanding of the targeted phenomenon, be it a normal or an anomalous one.

Acknowledgements We would like to thank all conference participants for their comments and suggestions on an early version of this chapter. Special thanks go to Kevin Hoover for his detail comments. Chao's research was supported by the Taiwan National Science Council (grant no. NSC100-2628-H-007-023-MY2).

References

Arbuthnot, John. 1710. An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London* 27: 186–190.

Blumberg, Baruch S. 2002. *Hepatitis B: The hunt for a killer virus*. Princeton: Princeton University Press.

Blumberg, Baruch S. 2006a. The curiosities of hepatitis B virus. *Proceedings of the American Thoracic Society* 3: 14–20.

Blumberg, Baruch S. 2006b. Hepatitis B virus conjectures on human interactions and origin of life. In *Life as we know It*, ed. Joseph Seckbach, 215–235. Dordrecht: Springer.

Blumberg, Baruch S., and Emily Oster. 2007. Hepatitis B and sex ratios at birth: Fathers or mothers? Working paper, The University of Chicago.

Cartwright, Nancy. 1997. What is a causal structure? In *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences*, ed. Vaughn R. McKim and Stephen P. Turner, 343–357. Notre Dame: University of Notre Dame Press.

Cartwright, Nancy. 1998. Capacities. In *The handbook of economic methodology*, ed. John B. Davis, D. Wade Hands, and Uskali Mäki. Cheltenham/Northampton: Edward Elgar.

Cartwright, Nancy. 1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.

Chahnazarian, Anouch. 1988. Determinants of the sex ratio at birth: Review of recent literature. *Social Biology* 35: 214–235.

Chen, Szu-Ting. 2011. Imagining the imaginable: A reinterpretation of the function of economists' concern about structural isomorphism in economic theorizing. *Journal of Economic Methodology* 18: 53–78.

Das Gupta, Monica. 2005. Explaining Asia's 'missing woman': A new look at the data. *Population and Development Review* 31: 529–535.

Das Gupta, Monica. 2006. Cultural versus biological factors in explaining Asia's 'missing women': Response to oster. *Population and Development Review* 32: 328–332.

[6] For the study of causal images, see Chen (2011).

Drew, Jean S., W. Thomas London, Edward D. Lustbader, Jana E. Hesser, and Baruch S. Blumberg. 1978. Hepatitis B virus and sex ratio of offspring. *Science* 201: 687–692.

Graunt, John. 1662. *Natural and political observations made upon the bills of mortality*. London: Martyn.

Haavelmo, Trygve M. 1944. The probability approach in econometrics. *Econometrica* 12: 1–115.

Hacking, Ian. 2006. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*, 2nd ed. Cambridge: Cambridge University Press.

Hempel, Carl. 1966. *The philosophy of natural science*. Englewood Cliffs: Prentice-Hall.

Hesser, Jana, Ionna Economidou, and Baruch S. Blumberg. 1975. Hepatitis B surface antigen (Australia antigen) in parents and sex ratio of offspring in a Greek population. *Human Biology* 47: 415–425.

Lahart, Justin. 2008. Economist scraps hepatitis theory on China's 'missing women'. *The Wall Street Journal*, May 22.

Lin, Ming-Jen, and Ming-Ching Luoh. 2008. Can hepatitis B mothers account for the number of missing women? Evidence from three million newborns in Taiwan. *American Economic Review* 98: 2259–2273.

Lipton, Peter. 2000. Inference to the best explanation. In *A companion to the philosophy of science*, ed. W.H. Newton-Smith, 184–193. Oxford: Blackwell.

Lipton, Peter. 2005. *Inference to the best explanation*, 2nd ed. New York: Routledge.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67: 1–25.

Oster, Emily. 2005. Hepatitis B and the case of the missing women. *Journal of Political Economy* 113: 1184–1186.

Oster, Emily. 2006. On explaining Asia's 'missing women': Comment on Das Gupta. *Population and Development Review* 32: 323–327.

Oster, Emily, Chen Gang, Yu Xinsen, and Lin Wenyao. 2008. Hepatitis B does not explain male-biased sex ratios in China. NBER working paper 13971.

Oster, Emily, Chen Gang, Yu Xinsen, and Lin Wenyao. 2010. Hepatitis B does not explain male-biased sex ratios in China. *Economics Letters* 2: 142–144.

Qian, Nancy. 2008. Missing women and the price of tea in China: The effect of sex-specific earnings on sex imbalance. *Quarterly Journal of Economics* 123: 251–1285.

Sen, Amartya. 1990. More than 100 million women are missing. *New York Review of Books*, December 20.

Sober, Elliot. 2007. Sex ratio theory, ancient and modern: An eighteenth-century debate about intelligent design and the development of models in evolutionary biology. In *Genesis redux*, ed. Jessica Riskin, 131–162. Chicago: The University of Chicago Press.

Steel, Daniel. 2008. *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

Steel, Daniel. 2013. Mechanisms and extrapolation in the abortion-crime controversy. In *Mechanism and causality in biology and economics*, ed. Hsiang-Ke Chao, Szu-Ting Chen, and Roberta L. Millstein, 185–206. Dordrecht: Springer.

Author Biographies

Marcel Boumans is an associate professor of history and methodology of economics both at the University of Amsterdam and Erasmus University Rotterdam and fellow of the Tinbergen Institute. He is coeditor of the *Journal of the History of Economic Thought*. His research is marked by three Ms: modeling, measurement, and mathematics. His main research focus is on understanding empirical research practices in economics from (combined) historical and philosophical perspectives. On these topics, he has published a monograph, *How Economists Model the World into Numbers* (Routledge, 2005), edited the volume *Measurement in Economics: A Handbook* (Elsevier, 2007), coauthored a textbook *Economic Methodology: Understanding Economics as a Science* (Palgrave-Macmillan, 2010), and coedited the *HOPE* annual supplement *Histories on Econometrics* (Duke University Press, 2011).

Hsiang-Ke Chao is an associate professor of economics at National Tsing Hua University, Taiwan. He received his doctoral degree in economics at the University of Amsterdam in 2002 and has been a visiting scholar at the London School of Economics, the University of California at Davis, and Stanford University. His current research focuses on model-based reasoning in history and philosophy of science, with particular interest in economics. Among his publications is *Representation and Structure in Economics: The Methodology of Econometric Models of the Consumption Function* (Routledge, 2009).

Ruey-Lin Chen is a professor of philosophy of science at National Chung Cheng University, Taiwan. He is the author of three books in philosophy of science and science fiction, two textbooks in critical thinking and philosophy of science, and over 30 articles in philosophy of science and other fields, mainly in Chinese. His research has been published in such journals as *Studies in History and Philosophy of Science*, and *East Asian Science, Technology and Society: An International Journal*. He is the coeditor, with Chienkuo Michael Mi, of *Naturalized Epistemology and Philosophy of Science* (Rodopi, 2007).

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9,

Szu-Ting Chen is an associate professor at the Graduate Institute of Philosophy, National Tsing Hua University, Taiwan. After receiving his BS degree in finance and MA degree in economics from the University of Oregon, USA, he moved to the London School of Economics where, under the supervision of Nancy Cartwright, he completed a PhD thesis on a metatheoretical account of economic theorizing. His main research interests include causation and explanation, philosophical foundation of economic methodology, philosophy of social science, and philosophy of science, in general.

Carl F. Craver is a professor of philosophy and a member of the Philosophy-Neuroscience-Psychology Program at Washington University in St. Louis. He is the author of *Explaining the Brain* (Clarendon, OUP), which provides a defense of mechanistic explanation in neuroscience, and (with Lindley Darden) *The Search for Mechanisms* (Chicago UP). In addition to his numerous articles in the philosophy of science, Craver also has an active scientific research program concerning the possibility of agency and decision-making in persons with amnesia.

Lindley Darden is professor of philosophy and a distinguished scholar/teacher at the University of Maryland, College Park in the USA. Her books include *Reasoning in Biological Discoveries* (Cambridge University Press, 2006) and *Theory Change in Science* (Oxford University Press, 1991). Her paper with Peter Machamer and Carl F. Craver, "Thinking About Mechanisms" (*Philosophy of Science*, 2000) proposed the "MDC" characterization of biological mechanisms; she responded to various critiques in "Thinking Again About Mechanisms" (*Philosophy of Science*, 2008). She and Carl F. Craver will publish *In Search of Mechanisms: Discoveries Across the Life Sciences* (forthcoming 2013, University of Chicago Press). She is a consulting editor to *Studies in History and Philosophy of Biological and Biomedical Sciences* and serves on the Editorial Board of *Philosophy of Science*. Her current research is on the history of computational biology and computational methods for discovering biological mechanisms.

Till Grüne-Yanoff is an associate professor of philosophy at the Royal Institute of Technology, Stockholm. His research focuses on the philosophy of science and on decision theory. In particular, he investigates the practice of modeling in economics and other social sciences, develops formal models of preference change, and discusses the use of models in policy decision-making. Till is also a member of the TINT Finnish Centre of Excellence in the Philosophy of Social Science in Helsinki.

Kevin D. Hoover is professor of economics and philosophy at Duke University. He is the author of *Causality in Macroeconomics* and *The Methodology of Empirical Macroeconomics*. He is a past editor of the *Journal of Economic Methodology* and current editor of *History of Political Economy*.

Marie I. Kaiser is a postdoctoral research fellow in the DFG research group on "Causation and Explanation" at the University of Cologne, Germany. She is mainly interested in philosophy of biology, in particular in reductive explanation, mechanisms, causation, and complex systems. She is the author of "The Limits of Reductionism in the Life Sciences" (Hist Philos Life Sci.; 2011).

Roberta L. Millstein is professor of philosophy at the University of California, Davis. Her research interests include the way that general topics in the philosophy of science, such as causation, mechanisms, probability, and determinism, illuminate and are illuminated by topics in evolutionary biology and ecology. She has recently published on the concepts of "fitness," "population," and "random drift"; on sexual selection; and on connections between population genetics and ecology.

Julian Reiss is a professor of philosophy at Durham University and specializes in philosophy of economics and general philosophy of science. His publications include *Error in Economics: Towards a More Evidence-Based Methodology* (Routledge 2008), *Causality Between Metaphysics and Methodology* (forthcoming with Routledge), "Causation in the Sciences: An Inferentialist Account," *Studies in the History and Philosophy of Science C* (2012), "The Explanation Paradox," *Journal of Economic Methodology* **19**(1): 43–62 (2012), "In Favour of a Millian Proposal to Reform Biomedical Research," Synthese **177**(3): 427–47 (2010), "Causation in the Social Sciences: Evidence, Inference, and Purpose," *Philosophy of the Social Sciences* **39**(1): 20–40 (2009), and "The Philosophy of Simulation: Hot New Topic or Same Old Stew?" (with Roman Frigg), *Synthese* **169**(3): 2009.

Daniel Steel is an associate professor in the Department of Philosophy and a faculty member of the Environmental Science and Policy Program at Michigan State University. His research focuses on evidence and causal inference especially as these topics pertain to the social and biological sciences. Dr. Steel is the author of *Across the Boundaries: Extrapolation in Biology and Social Science* (2008) and coeditor with Francesco Guala of *The Philosophy of Social Science Reader* (2011). In addition, he is the author of numerous articles that have appeared in such journals as *Philosophy of Science*, *Philosophy of the Social Sciences*, *British Journal for the Philosophy of Science*, and *Biology and Philosophy*.

David Teira is an associate professor at the Department of Logic, History and Philosophy of Science (UNED, Madrid) and deputy editor of THEORIA. He has published on the uses of statistics in economics and medicine and currently leads a research project on the correction of biases in medical experiments. For further information, visit http://www.uned.es/personal/dteira/.

Rong-Lin Wang is an assistant professor of philosophy at National Taiwan University. He received his PhD in Philosophy from Université Paris-Sorbonne. He taught previously at National Tsing Hua University (Taiwan). His research interests include philosophy of science, philosophy of biology, and ethics. He is the author of *Réalisme et antiréalisme dans la philosophie des sciences contemporaine* (Atelier National de Reproduction des Thèses, 2010).

Name Index

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9,

Subject Index

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*, History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9,