universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

"Simulating Earthquakes:
Explaining and Understanding Natural Phenomena
with Highly Idealized Models"

verfasst von / submitted by

Hernán Felipe Bobadilla Rodríguez, M.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Philosophie (Dr.phil.)

Wien, 2020 / Vienna, 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 792 296
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Philosophie
Betreut von / Supervisor:	Univ.-Prof. Dr. Martin Kusch
	Univ.-Prof. Dr. Karl Sigmund

Abstract

Scientists often use highly idealized models to explain and understand natural phenomena. The philosophy of scientific modeling seeks to understand how this is possible. One problem is simply how models can foster genuine explanations and understanding of their targets *despite the fact that they differ significantly from these targets.* Moreover, depending on how these models are used, they can be more or less informative about the targets.

The general purpose of this dissertation is to provide components of a solution to these conundrums. I develop an argument for the genuinely explanatory role of highly idealized models and their capacity to advance the understanding of natural phenomena. I craft and test this argument in the context of two highly idealized models of earthquakes which serve as case studies: i) the Burridge-Knopoff "spring-block" model of earthquakes and ii) the Olami-Feder-Christensen "cellular automaton" model of earthquakes.

The argument of this dissertation is as follows. In order to be used for model explanations, scientists must provide their models with more or less definite interpretations. I suggest that these interpretations are guided by scientists' explanatory commitments. As a result, the content of the ensuing model explanations and their epistemic status are relative to the interpretation embodied in the model. This relativity also applies to the idealizations in the model. To deal with these idealizations, I suggest prefixing model explanations with modal qualifications. This way, model explanations based on highly idealized models can be imputed to the target as a how-possibly, how-plausibly, and how-actually model explanation. It is worth noting that how-possibly and how-plausibly explanations provide genuine insight about the actual world, namely by stating possible or plausible ways to bring about the kind of phenomenon that the actual target instantiates.

Concerning scientific understanding, I suggest that genuine model explanations with modal qualifications advance scientific understanding in the form of explanatory understanding. I make this claim in the context of a newly developed framework for scientific understanding with models. In this framework, I distinguish explanatory and objectual understanding and consider the former to be a special case of the latter. I also argue that scientists who possess the same type of understanding – whether explanatory or objectual – may possess different states of understanding based on the content of their understanding. This content is relative to the interpretations that are embodied in the models. I suggest that these different states of understanding are valuable for scientific research. In particular, I argue that understanding differently plays an exploratory role which may be materialized in two distinct modes, namely the programmatic and prospective modes of exploration.

The main contributions of this thesis are: i) an account of genuine model explanation with modal qualifications; ii) a commitment-based account of interpretation in models; and iii) a new account of scientific understanding with models which integrates explanatory and objectual understanding and serves as a broader framework that contains the previous two accounts. Secondary contributions of this dissertation are: i) the introduction of novel case studies in discussions of explanations and understanding with models, namely models of earthquakes; ii) the introduction of a new model of "scientific research programs" based on scientific commitments; iii) a critique of the efficacy of philosophical accounts of explanation in describing scientific explanatory practices; and iv) the introduction of a distinction between two modes of exploratory research with models, namely programmatic and prospective exploration.

Zusammenfassung

In den Wissenschaften werden häufig höchst idealisierte Modelle dazu verwendet Naturphänomene zu erklären und zu verstehen. Die Wissenschaftsphilosophie dieser Modellierungen versucht wiederum zu verstehen wie das überhaupt möglich ist. Eine Problemstellung hierbei ist, wie wissenschaftliche Modelle eine Erklärung und ein Verständnis ihrer Gegenstände herbeiführen können, angesichts der Tatsache, *dass Modelle und ihre Gegenstände sehr unterschiedlich sind*. Je nachdem wie Modelle anwendet werden, können sie zudem auch mehr oder weniger informativ sein was ihre Gegenstände betrifft.

Der allgemeine Zweck dieser Dissertation ist es Bestandteile einer Lösung dieser Rätsel bereitzustellen. Mit meinem Argument verteidige ich die tatsächlich erklärende Rolle höchst idealisierter Modelle und ihr Vermögen unser Verständnis ihrer Gegenstände zu erweitern. Ich entwickle und prüfe mein Argument im Rahmen zweier Fallstudien zu idealisierten Modellen von Erdbeben: i) das Burridge-Knopoff "spring-block" Modell und ii) das Olami-Feder-Christensen "cellular automaton" Modell.

Mein Argument ist das folgende. Modelle bedürfen mehr oder weniger bestimmter Interpretationen, um in wissenschaftlichen Erklärungen verwendbar zu sein. Diese Interpretationen werden dabei von bestimmten Erklärungszwecken der Wissenschaftlerinnen und Wissenschaftler geleitet. Der Inhalt von erklärenden Modellen und deren erkenntnistheoretischer Gehalt verhalten sich daher zugleich relativ zu den in den Modellen enthaltenen Interpretationen derselben. Diese Relativität gilt ebenso von den in Modellen vorkommenden Idealisierungen. Ich schlage vor wissenschaftlichen Idealisierungen modale Bedingungen voranzustellen und Modelle somit zu unterscheiden danach wie wahrscheinlich, wie plausibel und wie tatsächlich sie ihre jeweiligen Gegenstände erklären. Die Plausibilität und Wahrscheinlichkeit von Erklärungen betreffen demnach immer noch die tatsächlichen Beziehungen der Modelle zu ihren Gegenständen, indem wie plausibel und wie wahrscheinlich sie Mittel und Wege beschreiben die Art des Phänomens hervorzubringen, welche in ihren Gegenständen instanziiert ist.

Mit Blick auf ihre modalen Bedingungen kann auch gesehen werden, wie Modelle nicht nur wissenschaftliche Erklärungen, sondern auch wissenschaftliches Verständnis voranbringen. Ich entwickle für diese Behauptung ein neues theoretisches Rahmenwerk für wissenschaftliches Verständnis anhand von Modellen. Innerhalb dieses Rahmenwerkes unterscheide ich zwischen erklärendem und gegenständlichem Verständnis, wobei ersteres ein Sonderfall des letzteren ist. Auch behaupte ich, dass Wissenschaftlerinnen und Wissenschaftler mit derselben Art von Verständnis – erklärend oder gegenständlich – aufgrund des Inhaltes ihres Verständnisses sich dennoch in verschieden Stadien ihres Verständnisses wiederfinden können. Dieser Inhalt verhält sich ebenfalls relativ zu den in Modellen enthaltenen Interpretationen. Ich behaupte jedoch, dass diese verschiedenen Stadien des Verständnisses durchaus wertvoll sind für wissenschaftliche Forschung. Verschiedene Verständnisse schlagen sich entsprechend nieder in eher programmatischen und eher zukunftsorientierten Formen der wissenschaftlichen Erkundung eines Gegenstandes.

Die Hauptbeträge dieser Dissertation sind die folgenden: i) eine Darstellung von Erklärungen anhand von Modellen mit modalen Bedingungen; ii) eine Auffassung von Interpretationen innerhalb von Modellen im Lichte ihrer jeweiligen Erklärungszwecke; iii) ein neues theoretisches Rahmenwerk für wissenschaftliches Verstehen mit Modellen, welches modale Bedingungen, Erklärungszwecke als auch erklärendes und gegenständliches Verstehen integriert. Zusätzliche Beiträge sind: i) die Einführung neuer Fallstudien mit der Erforschung wissenschaftlicher Modelle von Erdbeben; ii) die Einführung eines neuen Modells wissenschaftlicher Forschungsprogramme anhand von Erklärungszwecken; iii) eine Kritik philosophischer Auffassungen von Erklärung in der Beschreibung erklärender Praktiken in den Wissenschaften; und iv) die Einführung einer Unterscheidung zwischen zwei Formen der Erkundung durch Modelle in der Forschung, nämlich der programmatischen und der zukunftsorientierten Form.

Contents

Preface

There is something quite troubling about research proposals, especially those by PhD candidates. Research proposals are supposed to convey – among other things – research goals, methods, potential impact, and timetables. And they must do so convincingly in order to be approved. Still, for most PhD students, their doctoral research is their first in-depth and long-term research project. Thus, PhD students find themselves in a peculiar situation. They are expected to be able to design a research project and foresee its developments without having had comparable research experiences in the past. Not to speak of their lack of expert theoretical knowledge concerning the research subject in question. As a result, PhD research proposals tend to be no more than educated and elaborated statements of interest.

Supervisors know about this predicament. They try to do their best to provide their PhD students with bibliography, technical guidance, and encouragement. This way, the resulting proposal is as realistic a starting point as it can get. Still, as research unfolds, PhD students often follow trajectories different from those foreseen in the original proposal. For the most part, supervisors do not really expect their PhD students to follow their proposals to the letter. In fact, learning how to deal with unexpected developments by introducing adequate modifications is a crucial skill that is expected to be acquired during the PhD years. Regular meetings and yearly reports are occasions in which supervisor and supervisee are supposed to discuss the emerging problems of the research project and decide amendments together.

Besides unexpected obstacles, new research interests may be discovered as the investigation is conducted. Indeed, one has to recall that PhD students are not experts when they begin their research. Therefore, it is to be expected that their interests and opinions are going to be shaped by their inquiry. And it is reasonable for PhD students to try to include these new interests into their dissertation. After all, the PhD dissertation shapes the students' careers by providing them with definite opportunities for future job applications and as a platform for networking.

This dissertation embodies all the previous considerations. The final product is a monograph on how scientists explain and understand natural phenomena by means of highly idealized models, based on a study of models of earthquakes. Yet, my PhD research proposal – entitled "Causation in Seismic Models" – envisioned this dissertation as a study on causation. As part of the original project, I had a central hypothesis. I presumed that the explanations that seismologists crafted with the aid of models could be adequately described by means of well-established accounts of causation in the philosophical literature. With this assumption in mind, the goal was to find the right accounts of causation which captured the explanatory practices of scientists working with different kinds of models. Having identified the right accounts of causation, I could have been able to assess their efficacy in providing a fertile metaphysical foundation for successful explanations. Furthermore, having conducted this assessment, I anticipated that I could provide some criticism and advice on the most promising underlying causal assumptions for the modeling of earthquakes.

The original project proved itself to be rather naïve, on more than one occasion. There were, at least, four major problems that needed to be addressed. First, it soon became clear that I could not possibly deliver a systematic assessment of causal assumptions embedded in different kinds of models of earthquakes across seismology. Not to speak of providing criticism and advice on their underlying causal assumptions. I needed to focus my study on a few cases and, instead of criticism and advice for the scientists, I decided to restrict myself to a philosophical analysis of their practices. This way, the methodology of case studies became more prominent and the rather ingenuous normative component of the proposal was left out of the picture. My choice was to focus on two models of earthquakes, which are an exceptional pair to establish a compared analysis: they share a common history and embody some similar modeling assumptions. However, they ultimately respond to distinct research motivations, which are reflected in their distinctive design.

Second, even in the narrower context of these two case studies, I realized that seismologists' explanatory practices were not limited to the crafting of causal explanations. Noncausal approaches – namely mathematical – were also relevant explanatory strategies. This insight prompted major reengineering. I decided to shift the central topic of the dissertation from causation to explanation, allowing me to deal with the noncausal

approaches. Thus, instead of describing seismologists' modeling assumptions in terms of philosophical accounts of causation, I focused on describing their explanatory practices in terms of philosophical accounts of explanation. As a result, the overall tone of the dissertation became more epistemological than metaphysical, which I now consider to be one of its most fortunate amendments.

Third, I recognized that the description of explanatory practices by means of philosophical accounts of explanation was less straightforward than I expected. Seismologists – and scientists in general – do not always express their modeling and explanatory motivations explicitly. And even if they do, they might do so in ways that do not unequivocally match the philosophers' accounts of explanation. Thus, I learnt that the descriptive component of my proposal was a deeply interpretative task. This insight did not dissuade me from engaging in descriptions of scientists' explanatory practices but forced me to reflect upon the role of interpretation and elaborate on it.

As a result, I decided to introduce the notion of "explanatory commitments" into my argument. This notion enabled me to deal with two distinct problems related to interpretation. On the one hand, the notion of "explanatory commitments" allowed me to provide a more nuanced description of scientists' explanatory practices. This is because commitments are construed as a more fine-grained entity than accounts of explanation. This was a significant upgrade, given that scientists' explanatory practices do not always fit nicely within a single available philosophical account of explanation. On the other hand, explanatory commitments guide the interpretations in which scientists themselves engage with respect to their models. I submit that models must embody more or less definite – even if local – interpretations in order to be used by scientists. That is, models could be interpreted differently by distinct groups of scientists. In fact, my case studies provide evidence of that. I suggest that commitments play a major role in shaping the content of those interpretations.

Fourth, it even became unclear whether seismologists in the case studies were attaining explanations of the behavior of earthquakes at all. Their putative explanations were based on the content of highly idealized models that could say little about actual earthquakes. In trying to elucidate this matter, I discovered the literature on "scientific understanding".

This immediately became a new research interest. In this literature, I found effective tools that allowed me to better characterize the epistemic achievements gained with idealized models.

Thus, most of the topics in this dissertation were not foreseen in my proposal. The four aforementioned amendments carry most of the definitive character of the completed dissertation. The two case studies lead the discussion. “Explanation” became the central subject. “Commitments” and “interpretation” became deeply intertwined with the description of scientists’ explanatory practices with models. And the notion of “scientific understanding” provides a theoretical synthesis for the whole argument.

Acknowledgements

There were moments in which staying motivated with an ever-changing dissertation was extremely difficult, especially considering some dramatic extra-academic developments in the last stages of my writing process. As a consequence, I do not take the completion of this dissertation for granted. Several persons and institutions supported me through these years (more than I can mention here), whether economically, technically, or emotionally. First and foremost, I am particularly thankful to my supervisor, Martin Kusch, who guided me wisely through the difficulties of my dissertation and supported me beyond the technical. I am also thankful for the overall support of the faculty and fellows of the DK program "The Sciences in Historical, Philosophical and Cultural Contexts" (Austrian Science Fund "FWF" W1228-G18). I greatly benefited from two academic visits to the Science and Technology Studies Department at University College London and the History and Philosophy of Science Department at the University of Pittsburgh. I would like to thank all the researchers whom I met in these wonderful institutions. In particular, I am indebted to Phyllis Illari and James Woodward, who sponsored my visits to London and Pittsburgh, respectively. I would also like to express my gratitude to the Konrad Lorenz Institute, its fellows, and particularly to its scientific director, Guido Caniglia, for accepting me as a writing-up fellow and supporting me during the last stages of my dissertation. I thank Dejan Makovec for translating the final abstract of this dissertation (and for his friendship). Finally, I would like to thank my family and friends for their continuous support in the form of inspiration, guidance, and love.

1 Introduction

1.1 What Is the Problem?

It is extremely difficult to say anything general and, at the same time, accurate about scientific models. This is due to the great variety of models used across numerous scientific disciplines in distinct ways for dissimilar purposes. Arguably, the least controversial claim that one could submit about scientific models is that they are a crucial tool in most present-day scientific research, especially in the natural sciences. Besides that, the most general and accurate thing to say about models may be that there is nothing general and accurate that could be said about them.

This is somehow problematic for philosophers of science. Traditionally, philosophers of science have attempted to craft general accounts of scientific practices. Scientific modeling is no exception. There is a myriad of general philosophical accounts in the literature on what models are, on how models represent, and on how they are used to explain phenomena. Given that philosophers typically display generalist motivations, it is perplexing to witness the vast number of different, and often incompatible, accounts available in the literature. To be fair, some of these general accounts partly succeed in describing reasonably common features of scientific models. And, even if they fail to describe accurately specific cases, these accounts still provide expedient frameworks which enable philosophers to approach the study of models, even if as a starting point.

The motivations behind the crafting of general accounts can be expressed, at least partly, as the attempt of providing general solutions to seemingly general philosophical problems. When it comes to models, there are several of these problems. At its core, this dissertation aims to be a contribution to mainly two general philosophical problems about models and their epistemic role in scientific research. The first problem can be referred to as the problem of "surrogative reasoning." The second problem may be characterized as the problem of the "pragmatics of models."

The problem of surrogative reasoning can be expressed as follows. Scientific models are typically used as a surrogate system to reason and hopefully learn about a phenomenon of interest. However, it is not clear whether scientists are justified in assuming that models provide any insights into the targeted phenomenon. After all, these models – quite simply – are not the target phenomenon. As a general solution, philosophers of science have suggested that models provide epistemic access to their target phenomena insofar as there is some adequate relation between model and target. Several philosophers characterize this adequate relation as one of resemblance between model and target. However, as a general solution, this proposal requires further elaboration. Especially considering that, in several cases, models seem to be quite unlike the targets to which they are supposed to provide epistemic access.

Even if models provide epistemic access to their targets, it is still not clear whether this access is based on an objective relation between model and target. If it is not, then one might easily surmise that, depending on how models are used, they may or may not provide insights into their targets. Or even more, depending on how they are used, they may provide different kinds of insight into their targets. This brings me to the problem of the pragmatics of models. Anyone acquainted with actual scientific practices knows that models are highly versatile objects. One model may be used in different ways, for dissimilar purposes, with slight or major modifications, to provide insights to distinct targets. Accordingly, one has to ask oneself how to appraise these insights provided by a model relative to its mode of employment. One could even ask whether there are correct – or more expedient – modes of employment for acquiring specific kinds of insight.

The problems of surrogative reasoning and the pragmatics of models are broad and complex enough to be treated in several dissertations. In this one, I only address special cases of these general problems. With regards to the problem of surrogative reasoning, I focus on studying one of its more common expressions, namely that of explaining phenomena with models. In this case, scientists use the resources available in models to craft explanations of phenomena. The resulting explanations are commonly referred to as “model explanations.” Given that models are typically unlike their targets, in more than one sense, the problem is that it is not clear what the epistemic status of the ensuing model explanations is. In other

words, it is not clear whether model explanations are genuinely explanatory. Deciding this issue is particularly important in the context of model explanations based on unrealistic or highly idealized models.

With regards to the problem of the pragmatics of models, I mostly focus on the issue of how models – in order to be used – are bestowed with meaning. I refer to this issue as the "interpretation in models." (Note that I do not use the expression "interpretation *of* models" for reasons that will become clear later.) I submit that interpretations in models are guided by scientists' "commitments." Given that scientists may hold different commitments, their interpretations in models are likely to differ too. Then, the problem is to decide how to appraise the plurality of dissimilar potential interpretations that may be embodied in models.

As the reader might have anticipated, these two problems – the interpretation of models and the obtaining of genuine model explanations – are interrelated. Their interrelations can be spelled out as mutual influences. On the one hand, the interpretation in a model influences the kind of model explanations that may be crafted with it. On the other hand, given that scientists are mostly in the business of crafting genuine model explanations, there may be restrictions – even if contextual ones – on the kind of interpretations that a model should embody. The study of these interrelations is a sort of meta-problem, which I also tackle in this dissertation. I refer to this problem as the issue of "scientific understanding with models."

In sum, this dissertation deals with five problems. Two of them are dealt with as general problems, namely i) surrogative reasoning and ii) the pragmatics of models. Another two problems are special cases of the aforementioned problems and receive a detailed treatment, namely iii) genuine model explanations and iv) interpretation in models. And there is a final meta-problem which frames the previous ones, namely v) scientific understanding with models.

1.2 Goals and Scope

The main goal of this dissertation is to contribute to the solution of the aforementioned philosophical problems. I do not envisage to solve these issues completely or definitely. Still, I do expect to provide components of a solution or at least insightful discussions that may point towards adequate solutions. Given that the five main problems are general philosophical problems, my proposed solutions take the form of general philosophical accounts. More explicitly, I provide: i) an account of genuine model explanation based on modal qualifications; ii) a commitment-based account of interpretation in models; and iii) a new account of scientific understanding with models which integrates explanatory and objectual understanding and serves as a larger framework that contains the previous two accounts.

In addition to these main goals, this dissertation achieves other more specific aims which can be taken as complementary to the main ones. The most relevant of these specific aims are the following: i) to introduce novel case studies in the discussions of surrogative reasoning and the pragmatics of models, namely models of earthquakes; ii) to introduce a new model of "scientific research programs" based on scientific commitments; iii) to question the efficacy of philosophical accounts of explanation in describing scientific explanatory practices; and iv) to introduce a distinction between two modes of exploratory research with models.

Hoping to have conveyed the goals of this dissertation clearly, I would like to take a few lines to discuss the scope of the proposed general philosophical accounts. For the most part, philosophers of science are in the business of crafting philosophical accounts of scientific practices and scientific knowledge. These accounts are – so to speak – the models with which philosophers investigate science. Some of these models describe scientific practices more or less accurately. Some philosophers would go beyond description and use their models to prescribe specific forms of practice to scientists.

My proposed accounts are intended to be descriptive, not normative. But, as far as descriptive accounts go, mine have a rather limited scope. The reason for this is that my accounts are constructed based on the methodology of case studies. That is, I inspect cases in which scientists engage with seemingly idealized models and then resort to the ensuing

insights to inform the construction of my accounts. More explicitly, I assess how scientists interpret these models, how they use the models for explaining, and what it means to gain understanding via the models in these contexts. Insofar as my accounts attempt to capture my assessment of the case studies, they can be characterized as descriptive of the latter.

The reader might think that there is a contradiction between the restricted descriptive scope of my accounts and their intended generality. This is not the case. Generality and scope are, at least in principle, orthogonal notions. Generality is a formal feature of the accounts which is decided based on whether they are cast in categorical terms. In contrast, scope is the number of cases to which the accounts may be applied. Thus, one can see that an account may be applied to a wide variety of cases (i.e., have broad scope), even if it is not cast in categorical terms. A simple act of extrapolation would do the work.[1]

In my case, the situation is as follows. I do resort to categorical terms in my accounts, terms such as 'scientists', 'models', 'explanations', 'commitments', and the like. In this sense, my accounts are general. But their descriptive scope is restricted to my case studies. Or, to put it more precisely, the case studies are the "proven" descriptive scope of my accounts. However, I conjecture that my accounts may be applicable in other cases, especially those that hold some kind of resemblance to my case studies. In fact, I test components of my accounts in the context of other cases, even though these efforts are mostly ancillary to the more detailed treatment of my own case studies. In this sense, the generality of my accounts should be taken as an invitation to test them in other cases. This way, the proven descriptive scope of the accounts may be extended.

1.3 Methodologies

1.3.1 Literature Review

[1] For an enlightening discussion and application of the distinction between scope and generality, see Sheredos (2016: 928-930).

A central method in this dissertation – and arguably in any philosophical investigation – is the literature review. It consists of a general survey of scholarly texts on topics relevant to the problems at hand. These texts range from book-length essays, edited collections and doctoral monographs to special issues in journals, single peer-reviewed articles, conference contributions, and even preprints.[2] In this dissertation, the literature review can be divided into two main fields, namely the philosophical and the scientific. Within each field, several topics are investigated which have their own more specific literature.

The relevant philosophical literature for this dissertation is broad and diverse. This makes it rather challenging to provide an overview of the literature that influences my views and the texts that are ultimately cited in this dissertation. Still, three main topics and associated bodies of literature outstand. I provide a brief overview of the central book-length texts on these topics. To begin with, there is the literature on *scientific models* and *scientific representation*. In a few decades, this literature has become overwhelmingly vast. In order to navigate this literature, handbooks, encyclopedia entries, edited collections, special issues, and monographs were extremely useful. Worth of special mention are: Lorenzo Magnani and Tommaso Bertolotti's 2017 "Springer Handbook of Model-based Science," Axel Gelfert's 2016 "How to Do Science with Models," Michael Weisberg's 2013 "Simulation and Similarity," Daniela Bailer-Jones' 2009 "Scientific Models in Philosophy of Science," Mauricio Suarez's 2008 "Fictions in Science," Magnani and Nancy Nerssessian's 2002 "Model-based Reasoning," Magnani, Nersessian, and Paul Thagard's 1999 "Model-based Reasoning in Scientific Discovery," Mary Morgan and Margaret Morrison's 1999 "Models as Mediators," and Roman Frigg and Stephan Hartmann's 2020 entry on "Models in Science" in the *Stanford Encyclopedia of Philosophy*.

Another important body of literature in this dissertation is the one on *scientific explanations*. Wesley Salmon's 1989 "Four Decades of Scientific Explanation" remains a great introduction to the subject. More recent overviews are James Woodward's 2019 entry on "Scientific Explanation" in the *Stanford Encyclopedia of Philosophy* and Eric Weber, Jeroen Van Bouwel, and Leen De Vreese's 2013 "Scientific Explanation" (even though the latter is

[2] For more details on "literature reviews," see Race (2008).

also a book with a specific thesis). Particularly influential accounts of explanation are studied in landmark essays and monographs, such as: Carl Hempel's 1965 "Aspects of Scientific Explanation," Philip Kitcher's 1989 "Explanatory Unification and the Causal Structure of the World," Bas van Frassen's 1980 "The Scientific Image" (particularly Chapter 5), Woodward's 2003 "Making Things Happen," and Michael Strevens' 2008 "Depth."

Within the literature on scientific explanation, I review texts on more specific discussions. In particular, I review the literature on new mechanist accounts of explanation. Particularly relevant in this regard are Stuart Glennan's 2017 "The New Mechanical Philosophy," Glennan and Phyllis Illari's 2017 "Routledge Handbook of Mechanisms and Mechanical Philosophy," William Bechtel and Robert Richardson's 2010 "Discovering Complexity," and Carl Craver's 2007 "Explaining the Brain." I also review noncausal accounts of explanation. Salient books on the topic are Alexander Reutlinger and Juha Saatsi's 2018 "Explanation Beyond Causation," Morrison's 2015 "Reconstructing Reality," Christopher Pincock's 2012 "Mathematics and Scientific Representation," and Robert Batterman's 2002 "The Devil in the Details." In addition, papers by Colin Rice, Marc Lange, Philip Huneman, and Daniel Kostic are relevant sources for discussions on noncausal accounts. Finally, Alisa Bokulich's various contributions on model explanations deserve a special mention, given their position at the interface between the literature on scientific models and scientific explanations.

Finally, I also explore the philosophical literature on *scientific understanding*. This is a comparatively recent literature, although with promising prospects. Particularly salient in my research are five monographs: i) Henk de Regt's 2017 "Understanding Scientific Understanding"; ii) Catherine Elgin's 2017 "True Enough"; iii) Kareem Khalifa's 2017 "Understanding, Explanation, and Scientific Knowledge"; iv) Angela Potochnik's 2017 "Idealization and the Aims of Science"; and v) Jonathan Kvanvig's 2003 "The Value of Knowledge and the Pursuit of Understanding," especially Chapters 8 and 9. Two edited collections are also relevant sources, namely Henk de Regt, Sabina Leonelli, and Kai Eigner's 2009 "Scientific Understanding" and Stephen Grimm, Christoph Baumberger, and Sabine Ammon's 2017 "Explaining Understanding."

The scientific literature review is mostly shaped by the *case studies*. In this sense, there are two central papers which introduce the cases to be studied, namely Robert Burridge and Leon Knopoff's 1967 "Model and Theoretical Seismicity" (hereinafter "BK") and Zeev Olami, Hans Jacob Feder and Kim Christensen's 1992 "Self-Organized Criticality in a Continuous, Nonconservative Cellular Automaton Modeling Earthquakes" (hereinafter "OFC"). I also review additional publications by the same authors. This strategy is particularly relevant in the case of OFC, given that they published their simulation results in a series of papers. I also review secondary literature in which the BK and OFC models are discussed. In addition, I review literature on the subject of "Self-organized Criticality" (hereinafter "SOC"), which is relevant to fully comprehend OFC's contribution. Particularly salient are four book-length texts on the subject: i) Per Bak's 1996 "How Nature Works"; ii) Henrik Jensen's 1998 "Self-Organized Criticality"; iii) Stefan Hergarten's 2002 "Self-Organized Criticality in Earth Systems"; and iv) Gunnar Pruessner's 2012 "Self-Organised Criticality." In addition, a particularly useful source is a special issue of Space Science Reviews 2016 (vol. 198) which collects a series of state-of-the-art papers, in commemoration of the 25 years of SOC. Worth of a special mention is Frigg's 2003 paper "Self-organised criticality – what it is and what it isn't," as one of the few philosophical explorations of SOC. I also resort to more general literature on earthquake models and geophysics. I highlight Yan Kagan's 2014 "Earthquakes" and Donald Turcotte's 1997 "Fractals and Chaos in Geology and Geophysics."

1.3.2 Case Studies

Case studies are a qualitative method of research which tends to focus on descriptive and interpretive concerns. Case studies are usually single entities or contexts that instantiate a phenomenon of interest for the researcher. Blatter (2008) suggests that case studies can be used in two distinctive fashions. First, case studies can be used to test the validity and scope of theories. Such testing is admissible because the case studies (presumably) instantiate the phenomena that the theories address. Second, case studies can be used as a generative source for new theories. This is possible because case studies are well suited for description,

collection of evidence, and further explorative tasks. These two fashions – call them the "testing" and "generative" uses – should be seen as endpoints of a continuum.[3]

As a qualitative method in the philosophy of science, case studies exhibit advantages and disadvantages. Just to give a taste of some of these, consider the context in which case studies are used for testing theories. Arguably, the most salient advantage of case studies is the "depth" of analysis, as opposed to the "breadth" of other methods (e.g., large surveys and their respective statistical analysis). By depth, I mean that case studies allow for a detailed and nuanced assessment of theories. Such assessment contributes, among other things, to the achievement of consistency within the tested theories, consistency with other theories, and improvement of the overall conceptual apparatus. However, a comparative disadvantage of employing case studies for testing theories is the questionable external validity of the insights gained from them. After all, case studies might exhibit idiosyncratic features that do not reflect the general merits of a theory. In this sense, the researcher must be able to locate her case study in relation to other relevant contexts and assess how the case study informs these other contexts.[4]

A middle ground between testing and generating theories is often met whenever case studies are employed to inform the modification of existing theories. Felicitous case studies allow for the evaluation and development of more than one theory – or part of a theory – at a time. In fact, a comparative advantage of case studies over other methods is that they allow for the parallel testing of several aspects of a theory – or even several theories – at once. Thus, case studies are well suited to explore and describe the interrelations between distinct theories or distinct parts of a single theory. In this dissertation, I use the methodology of case studies in this way: a middle point between testing existing philosophical accounts, modifying some of their aspects, and eventually generating new ones.

In this dissertation, I focus on two case studies, namely the models of earthquakes presented by BK and OFC in their respective papers (hereinafter the "BK model" and "OFC

[3] For more details on "case studies," see Gerring (2007).
[4] For more details on the employment of case studies in the philosophy of science, see Curie (2015).

model," respectively). There are several reasons that justify choosing these models as case studies. To begin with, they fulfill the basic requirement of instantiating my subject matter. That is, they are highly idealized models used for explaining and understanding – or so I argue. In this sense, the BK and OFC models are well suited to test existing accounts of explanation and understanding with idealized models. Furthermore, specific modifications to these accounts can be explored in these case studies and, eventually, I postulate new accounts based on my study of them.

Second, the BK and OFC models – and the original papers in which they were introduced – have had a great and lasting impact. Such impact is manifested, for example, in the number of citations the original papers have received, some of them fairly recently. To make this more explicit, BK have been cited in 1603 papers, according to Google Scholar (8 August 2019) and 937 papers, according to Web of Science (8 August 2019), from which 16 are 2019's papers. And OFC has been cited in 1197 papers, according to Google Scholar (8 August 2019) and 768 papers, according to Web of Science (8 August 2019), from which 17 are 2019's papers. Beyond the number of citations, the impact of the BK and OFC models is also attested by their status as "standards" in textbooks and other scientific papers. For example, Pruessner (2012) uses the OFC model – and collaterally the BK model – as exemplars to discuss self-organized criticality in his textbook on the subject. To be sure, Pruessner discusses other related models as well. Notable examples are Carlson and Langer (1989), Nakanishi (1990) and Brown *et al.* (1991). However, Pruessner argues that the OFC model "finally generalized the [BK] Model comprehensibly and formulated it most succinctly" (126). In this sense, the OFC and BK models have become standards.

Third, the BK and OFC models are closely related to each other. To be more specific, the OFC model is an explicit extension of the ideas tested in the BK model (see OFC: 1244). Because of this, there are several similarities between the models, but also significant differences. More explicitly, the models are based on the same basic intuition, but they embody different implementations and representational assumptions. As a result of these relations, the insights reached in each case study are comparable and complementary. By combining these insights, I find myself in an expedient position to craft philosophical accounts with a broader scope but also nuanced.

Fourth, I would argue that there is an intrinsic value in using these case studies. This claim can be defended in two ways. The first way is that, by studying these models, I close a research gap in the philosophy of science. Indeed, the BK and OFC models have been neglected by philosophers of science. In addition, their target phenomena – namely earthquakes – have received scarce attention, not to mention the disciplines of seismology and geology in general. The second intrinsic value of these case studies is their impact on human lives. After all, these models are used to study a natural hazard which affects several communities across the globe. To put it bluntly, explanations and understanding of earthquakes are relevant not only for scientific and philosophical purposes. They are also relevant for more practical endeavors, from urban planning to mitigation policies. In this sense, philosophical reflection upon the science embedded in these case studies and the understanding that they afford have actual impact.

1.3.3 Content Analysis

Case studies can be conducted in various dissimilar ways. In my case studies, I mostly rely on a method that has been referred to as "content analysis." This method basically consists in "categorizing qualitative textual data into clusters of similar entities, or conceptual categories, to identify patterns and relationships between variables or themes" (Julien, 2008: 120). In other words, content analysis can be described as "making sense" of a text in light of a framing theory. In my dissertation, the relevant framing theories are various philosophical accounts of scientific models, scientific explanations, and scientific understanding.

In practice, this means that I engage in close reading of the original papers in which the BK and OFC models were first introduced. The purpose of such close reading is to identify the main themes that recur in the texts. In particular, I attempt to identify, describe, and assess three main topics: i) scientists' epistemic motivations and explanatory commitments; ii) the representational features and inner workings of the models, as expressed by the scientists; and iii) the role of the models in addressing the motivations of the scientists.

Content analysis is an openly interpretive method. There are more or less legitimate – or at least commonly accepted – ways of dealing with the interpretive component of content analysis. Two considerations are important. First, what is typically regarded as legitimate interpretations are constrained by intersubjective codes and conventions. Taking this into account, I moderate my interpretations in light of the interpretations embodied in the secondary literature on the BK and OFC models. Second, scientists and philosophers often use the same terms in their texts, but with different meanings. Because of this, content analysis is not synonymous with textual analysis. Content analysis must go beyond mere textual analysis. In other words, interpretations of content need to be contextualized. I take this into account and strive for caution distinguishing the philosophical from the scientific contexts. For example, attention must be drawn to the time of the publications, their respective journals, disciplinary idiosyncrasies, and even historically sound influences between philosophers and scientists.

1.4 Structure of the Dissertation

The structure of this dissertation is as follows. Chapter 1 is the introduction to this dissertation. In it, I convey: i) the main problems to be addressed; ii) the goals and scope of the dissertation; iii) the methodologies to be followed; and iv) the structure of the argument as presented in the dissertation.

In Chapter 2, I proceed to introduce the case studies with which I work for the rest of the dissertation, namely the BK and OFC models. The reason for this early introduction of the case studies is that subsequent chapters address key elements for the construction of my argument which are tested in the context of my case studies. In other words, the case studies enable the construction of my argument throughout the dissertation. In addition to the case studies, Chapter 2 also introduces the basics of earthquake occurrence in geological faults and the main tenets of a research program which is relevant to my case studies, namely Self-organized Criticality.

From Chapter 3 to Chapter 8, I construct my philosophical argument. Each chapter has its own inner structure which is presented at the beginning of the chapter. At the end of each

of these chapters, I provide a précis which contains the main theoretical theses presented in the chapter, together with the main lessons learnt in the context of my case studies.

In Chapter 3, I begin the construction of my philosophical argument by introducing the notion of "scientific commitments". Scientific commitments are guidelines that are accepted by scientists and shape their research. I submit that scientists who operate within research programs accept scientific commitments that are idiosyncratic to the program. Scientific commitments are organized within research programs in constellations of mutual support. I refer to these constellations as "accounts". Accounts play different roles in research programs and they typically exhibit a differential contribution to scientific endeavors conducted within the program. Because of this, I consider it crucial to discuss how commitments are justified in being part of accounts and research programs. In particular, I discuss a scheme of justification of commitments for programmatic contexts. In these contexts, justification can be epistemic or practical and it aligns with a coherentism-cum-entrenchment model. Basically, this means that commitments stand in mutual accord but they contribute differentially to that accord.

In Chapter 4, I focus on one kind of scientific commitment, namely explanatory commitments. I submit that explanatory commitments are organized in accounts of explanation which guide scientists' explanatory practices. Philosophers have tried to describe these accounts and have delivered what I refer to as philosophical accounts of explanation. In this sense, philosophical accounts of explanation are philosophers' articulation of scientists' accounts of explanation. For the most part, philosophical accounts of explanation have been used fruitfully to describe scientists' explanatory practices. Still, there are cases in which scientists' explanatory commitments may not be wholly captured by well-established philosophical accounts of explanation. I show this in the context of my case studies. After having explored and discussed these problems, I decide to keep using philosophers' accounts of explanation in my argument, but I do this with awareness of their limited capacities.

In Chapter 5, I argue that scientists' commitments play a crucial role in the employment of models, namely an interpretive role. Scientists must provide models with more or less definite interpretations in order for them to be used. I submit that scientists'

commitments guide these interpretations. In particular, if models are to be used for model explanations, explanatory commitments play a central interpretive role. Because interpretations are guided by scientists' commitments, they often differ from scientist to scientist, especially if they operate within distinct programs. As a result, whenever models are used for model explanation, the content of the ensuing model explanations is relative to the interpretation in the employed model. Furthermore, one or another interpretation prompts model explanations that are often assessed as having different epistemic status. In simple words, some interpretations are considered better than others.

This brings me to the topic of genuine model explanations in Chapter 6. For the most part, scientists aim for genuine model explanations, i.e., model explanations that provide genuine insight into the way target phenomena occur. However, given that the interpretations embodied in models often pose idealizations of various sorts, it is not clear how model explanations can be genuinely explanatory of their targets. In order to solve this predicament, I suggest that model explanations should be prefixed with modal qualifications. This way, the model explanation can be imputed to the target as a how-possibly, how-plausibly and how-actually model explanation. It is worth noticing that how-possibly and how-plausibly explanations provide genuine insight about the actual world, namely by stating possible or plausible ways to bring about the kind of phenomenon that the actual target instantiates.

Up to this point, a complete argument has been presented on how highly idealized models are used for model explanations. In Chapter 7, I reframe most of my argument in terms of a broader framework, which I introduce as my account of scientific understanding with models. This account distinguishes two types of understanding often found in the literature, namely explanatory and objectual understanding. I suggest that genuine model explanations with modal qualifications advance scientific understanding in the form of explanatory understanding. But I also suggest that understanding may go beyond the attainment of genuine model explanations, namely in the form of objectual understanding. In fact, I submit that explanatory understanding is only a special case of objectual understanding.

Finally, in Chapter 8, I argue that scientists who possess the same type of understanding – i.e., explanatory or objectual – may possess different states of understanding based on the content of their understanding. This content is relative to the interpretations that are embodied in the models that afford understanding to the scientists. I suggest that understanding differently is an important feature of scientific research. In particular, I argue that it plays an exploratory role that may be materialized in two distinct modes of exploration, namely the programmatic and prospective modes.

***

2 Case Studies: the BK Model and the OFC Model

My case studies amount to two scientific models which simulate aspects of the behavior of earthquakes produced at geological faults. The first one is the Burridge-Knopoff (BK) model (1967). The BK model is implemented as a laboratory model, i.e., a concrete material object, and as a physico-mathematical model, i.e., a set of differential equations solved via numerical methods implemented in a computer program. The second case study is the Olami-Feder-Christensen (OFC) model (1992), which is a cellular-automaton simulation implemented in a computer program. I present these models as they were first introduced in their corresponding seminal papers. I complement the presentation and discussion of these papers with further commentaries from secondary literature.

I go into some detail concerning the presentation of these models. I avoid technicisms, but I do not avoid presenting a great part of the content of the original papers. I do this now because I intend to use this content in the following chapters in order to build my argument and construct my philosophical accounts. In this sense, this chapter should be taken by the reader as a first opportunity to gain familiarity with the most significant elements of the case studies, which will recur along most of the dissertation. Although the intention of this chapter is mostly expository, I do engage in minimal discussion with regards to features of the models and seeming motivations of the respective researchers.

The structure of the chapter is as follows. I begin with a preliminary section on general features of earthquakes. This section is meant to set the background for the discussion. In this sense, it is not intended to be an exhaustive nor detailed review of earthquake theories. Then, I proceed to tackle my first case study, namely the BK model. I focus on four main issues. First, I briefly review the scientific context in which the BK model was introduced. This amounts to describing its relations to traditional seismology and qualitative models available at the time. Second, I present and discuss one of the implementations of the BK model, namely the laboratory implementation. Third, I present and discuss a second implementation of the BK model which amounts to a physico-mathematical description and its numerical solution via a

computer program. Fourth, I present the main results of the BK model. Finally, I close this chapter with the study of the OFC model. But, in order to do this, I must first go through three preliminary sections. The purpose of these preliminary sections is to introduce a research program – namely self-organized criticality (SOC) – which motivates the OFC model. Then, I proceed to describe in detail the implementation of the OFC model and the main results derived from simulations with it.

2.1 Preliminaries on Earthquakes

In its broadest sense, earthquakes are – as their name suggests – vibrations that propagate across the Earth. These vibrations have different origins. Some earthquakes are caused by human activities. For example, nuclear weapons tests, rocket launches, and even public transportation can cause vibrations on the Earth. Other earthquakes are caused by non-anthropogenic causes. For example, volcanic eruptions, landslides, and meteorite falls can produce earthquakes. In this dissertation, I focus on non-anthropogenic earthquakes of a particular kind, namely tectonic earthquakes.

To understand the processes that lead to tectonic earthquakes, it is crucial to grasp the basics of plate tectonics. According to this theory, the Earth's outer layer – known as the lithosphere – is fragmented in large plates, which behave rigidly over geological time (see Figure 2-1). These plates move relative to one another due to convection cells in the Earth's mantle below them. These relative motions among plates embody tectonic forces which build stress fields at the plates' boundaries. As a consequence of these stress fields, rocks at the boundaries of tectonic plates deform. For the purposes of discussing my case studies, I focus on deformation regimes that are typical of shallow portions of the lithosphere. In these shallow portions, rocks exhibit a mostly elastic and brittle behavior. This basically means that rocks deform by storing elastic energy up to a rupture threshold after which the stored energy is radiated in the form of a mechanical wave. This wave is a tectonic earthquake.

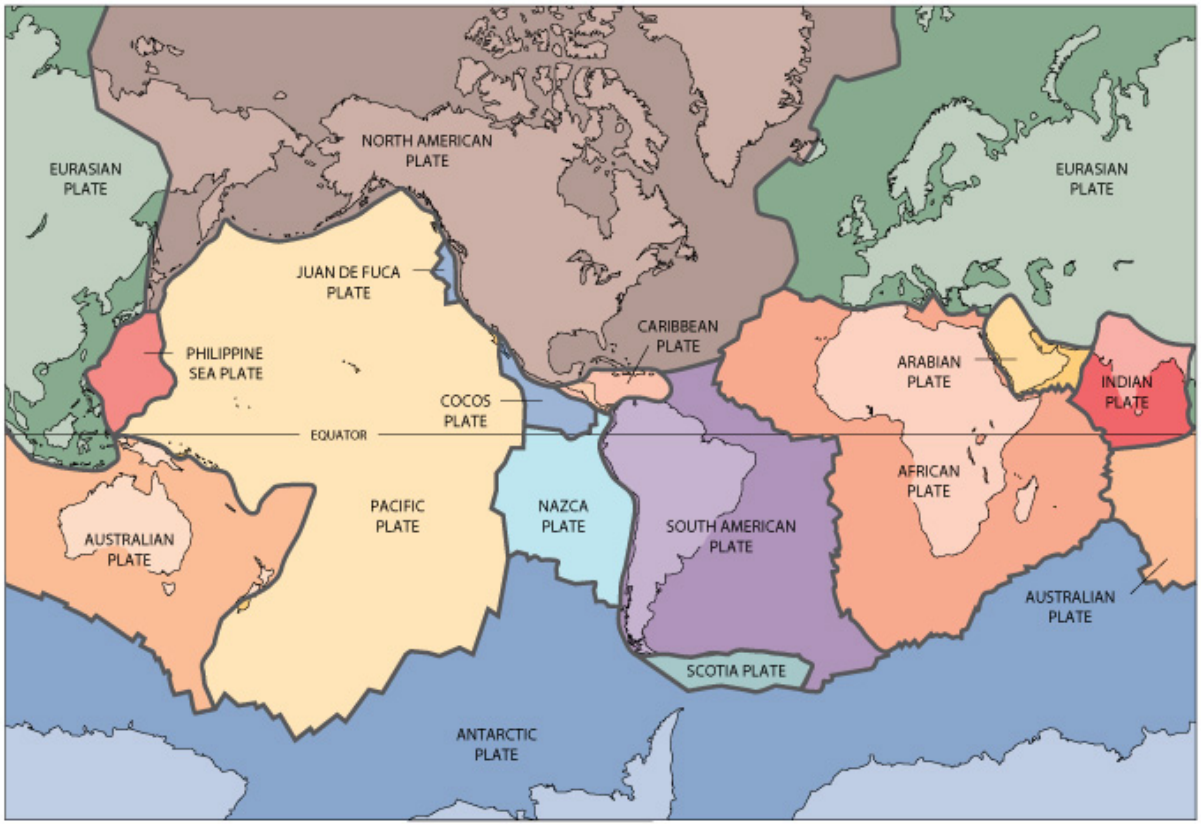


Figure 2-1: Major tectonic plates. Retrieved from US Geological Survey website. (http://pubs.usgs.gov/publications/text/slabs.html)

The elastic deformation and subsequent rupture that lead to tectonic earthquakes can be described in terms of a "stick-slip" mechanism. According to this mechanism, earthquakes occur cyclically in pre-existing geological faults, which are approximately planar discontinuities in rocks (see Figure 2-2). The masses of rocks on both sides of the fault are often referred to as "walls." The walls are affected by differential stresses – i.e., internal forces between different parts of the rock medium – mostly due to tectonic forces. These differential stresses prompt the material on both sides of a geological fault to slide relative to each other. However, frictional forces at the fault prevent such slide from happening. This is the "stick" component of the stick-slip mechanism. Consequently, stress builds up.[5] Eventually, the accumulated stress exceeds the static frictional forces at the fault. At this point, the masses of rock on both sides of the fault slide relative to each other. This is the "slip" component of

[5] Together with the building up of stress, there is an increase in strain. Roughly, strain is a measure of the deformation of three-dimensional bodies, such as masses of rock. Although stress and strain are conceptually different, they do not exist independently of each other. They are related in terms of the constitutive properties of a solid medium, namely its elasticity.

the stick-slip mechanism. Along with this sudden displacement, there is a release of stress, which propagates across the rocks as a seismic wave. Frictional forces at the fault stop the displacement of the rock walls along the fault and stress starts building up for the next cycle. For more details on the stick-slip mechanism, see the landmark paper by Brace and Byerlee (1966).

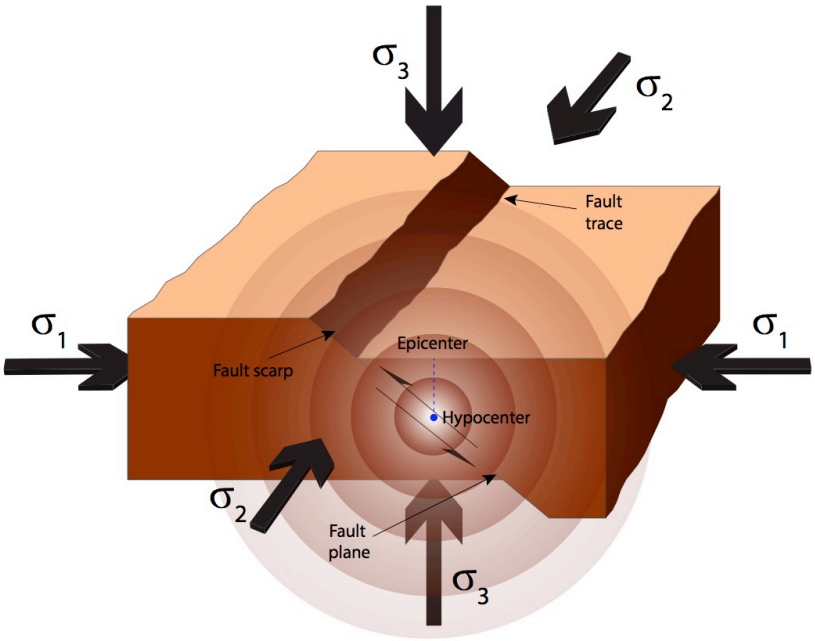


Figure 2-2: Scheme of a geological fault. The geological fault defines two walls, which are the contiguous blocks of rock. The black arrows represent three orthogonal axes of normal stress acting upon the rocks (σ_1, σ_2, and σ_3). An earthquake is represented as a wave which propagates in all directions. The earthquake originates at the fault, at the hypocenter. The projection of the hypocenter to the Earth's surface is referred to as the epicenter.

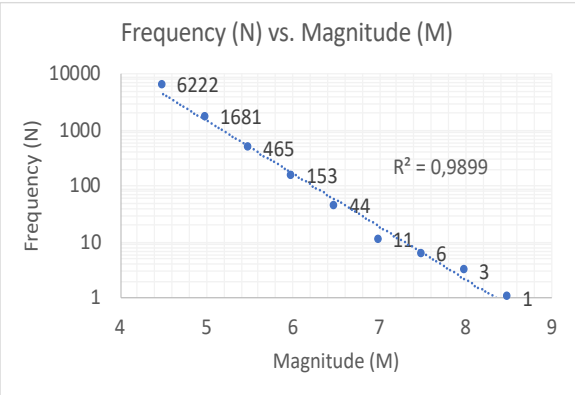
The occurrence of tectonic earthquakes – hereinafter simply "earthquakes" – follows several well-documented patterns of behavior in space and time. One of these patterns is particularly important for my discussion of the case studies. Across different seismic regions, and for arbitrary time intervals, the relation between earthquakes' magnitude and their number follows a robust pattern. This pattern has the status of an empirical statistical law and it is better known as the Gutenberg-Richter (GR) law. According to this pattern, the magnitude of earthquakes relates to their frequency in accordance with the following equation:

$$\log_{10} N = a - bM$$

where *N* is the number of earthquakes with magnitudes greater than a certain magnitude *M*, *a* and *b* are constants that vary across seismic regions. Given that an earthquake's magnitude is proportional to the area of rupture at a geological fault, the GR law can be characterized as a spatial pattern in time. As an illustration, consider the following case: for an arbitrary area at the Pacific border of South America, for an arbitrary time interval which covers the last 15 years, the data exhibits a statistically significant exponential regime (see Figure 2-3).

Area
Latitude: [-56.462, -16.13]
Longitude: [-77.915, 63.281]
Time interval
Starting date: 17.06.2005
Ending date: 24.06.2020

Magnitude (M)	Frequency (N)
8,5	1
8	3
7,5	6
7	11
6,5	44
6	153
5,5	465
5	1681
4,5	6222



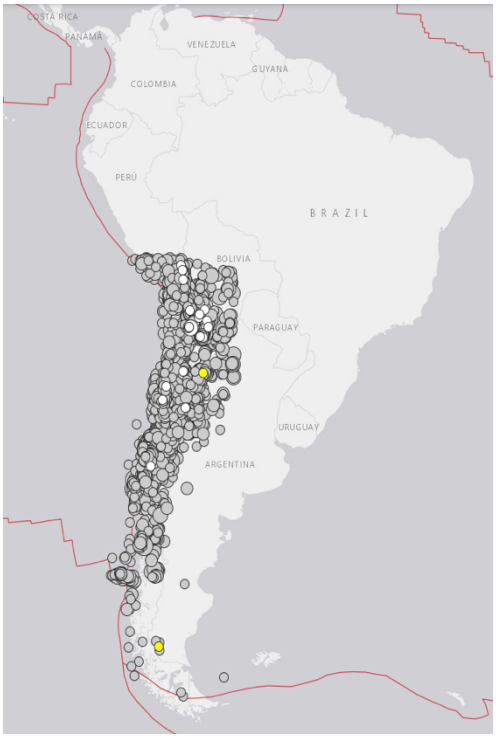


Figure 2-3: Data of earthquakes for arbitrary area and time interval. Data retrieved from the "Search Earthquake Catalog" (retrieved from "US Geological Survey" website, June 24th 2020).

The GR law can be presented in a slightly different way, which affords a more direct physical reading. The magnitude M of an earthquake is a function of the energy released by the earthquake according to the following relation:

$$M = \alpha + \beta \log_{10} E$$

Combining both equations:

$$\log_{10} \frac{N}{N_0} = -b\beta \log_{10} E$$

$$N = N_0 E^{-b\beta}$$

In other words, the relation between the number of earthquakes and their energy release follows a power-law distribution.

Seismologists have built several models to investigate the robustness and causes of the GR law. Among these models, my case studies are considered major landmarks. They simulate the power-law behavior of earthquakes based on a simple mechanism which mimics the stick-slip behavior at geological faults. In the following sections, I introduce and begin discussing these models. I proceed first with the BK model.

2.2 Case Study I: the Burridge-Knopoff (BK) Model

2.2.1 Scientific Context

Burridge and Knopoff's 1967 article – "Model and Theoretical Seismicity" – begins by introducing and distinguishing the notions of "theoretical" and "model" seismology. BK characterize both theoretical and model seismology in terms of their "well-known" principles and aims. On the one hand, theoretical seismology's main aim is "to solve the elastic wave equation for a given set of initial conditions and for a given configuration of geometrical boundaries" (BK: 341). This aim is characterized as a problem in "applied mathematics" (ibid). BK suggest that this problem is of a deductive nature, since "the partial differential equations of elastic wave motions are well-known and the geometry and the natural constraints of the prototype are presumed to be given" (ibid). On the other hand, model seismology aims for constructing "the most obvious analog computer in the laboratory for the corresponding problem of theoretical seismology" (ibid). In this sense, model seismology is derived from theoretical seismology: "The problem of model seismology is expressible as a *corollary* to that of theoretical seismology" (ibid; my emphasis).

It is worth clarifying the distinction between the notion of "seismology" – as used by BK – and BK's main concern, namely "seismicity." On the one hand, seismology – both theoretical and model – is mostly concerned with the study of the propagation of earthquakes across the Earth as seismic waves. Complementarily, seismologists also study topics that capitalize on knowledge of the propagation of earthquakes. For example, seismologists can

study the inner structure of the Earth, based on observations of how seismic waves propagate across the Earth. On the other hand, seismicity studies earthquakes but focuses on their occurrence, i.e., their causes, geographical distribution, frequency, and magnitude. The problem of occurrence of earthquakes is not part of seismology, as characterized by BK. However, BK's characterization of seismology has become rather outdated. Nowadays, it is broadly understood that the problem of the occurrence of earthquakes (i.e., seismicity) is a seismological concern.

BK suggest that a reason why seismicity is not studied in seismology is that there is no physical theory that predicts the occurrence of earthquakes in time and space. Or, to be more precise, there is no physical theory that describes the occurrence of earthquakes in analytic form as a function of temporal and spatial variables. Still, qualitative descriptions of earthquake occurrence had been delivered by the time BK published their paper, such as the stick-slip mechanism introduced above. The BK model is an attempt to simulate aspects of this mechanism. It is worth mentioning that BK do not use the term 'stick-slip'. However, as BK describe the mechanical assumptions of their model, it is clear that they are assuming a picture equivalent to that of the stick-slip mechanism. This is also supported by secondary literature (e.g., see Clancy & Corcoran, 2006: 046115-1).

The resulting model is basically a system of blocks connected through springs. This is why the BK model is often referred to as the "spring-block" model. The spring-block system is implemented in two ways.[6] First, an actual spring-block system is constructed as a concrete, material model in the laboratory. BK refer to this model as the "laboratory model." Two simulations with the laboratory model are reported in BK's paper. BK refer to these simulations as "experiments."[7] Second, the dynamics of a generalized version of the spring-block system is described in physico-mathematical terms. Such description is implemented –

[6] The distinction between a model and its distinct implementations resembles that of Mäki (2009b) between a model and its descriptions. For Mäki, a model is an imagined abstract system, which can be described in terms of different concrete schemes or materials (34). In this sense, the two implementations of the BK model can be referred to as two distinct model descriptions of the same imagined spring-block system.

[7] The overlap between experimentation and modeling is also noted by Mäki (2009a). He asserts that "models are experiments and experiments are models" (80). Such slogan addresses both material and theoretical models. Material models are experiments in the traditional sense of being manipulatable constructed systems which provide indirect epistemic access to a target phenomenon. And theoretical models can be taken as "thought" experiments in which the relations between theoretical principles and specific factors are tested.

that is, solved – as a numerical-computer model. BK refer to this model as the "numerical model." I prefer the label "mathematical implementation" because this implementation is not limited to the numerical solution of a physico-mathematical description. Rather, this implementation includes the physico-mathematical description. To put it differently, this implementation captures the theoretical and model approaches prevalent in seismology. The theoretical approach is embodied in BK's attempt to provide a physico-mathematical description of the problem at hand. And the model approach is embodied in the crafting of a computer program that solves the physico-mathematical description via a numerical method. (It is not an accident that BK decide to call their paper "Model and Theoretical Seismicity.")

Several simulations are run in both implementations, testing different settings. Varied results are then reported and discussed. The discussions are mostly shaped by comparisons between the simulation results and empirical statistical evidence from naturally occurring earthquakes. From all the results and discussions, I mostly focus on those that replicate a power-law between frequency and size of the model earthquakes. A main reason why I focus on these results is that they afford a convenient criterion to establish comparisons between the BK model and my second case study, namely the OFC model. Now, I proceed to introduce the BK model in its laboratory implementation.

2.2.2 BK Model – Laboratory Implementation

As a prelude to the description of the laboratory model, BK discuss another model, one that they did not actually construct. This other model may be regarded as a thought experiment whose purpose is adjusting some intuitions about the problem of earthquake occurrence. The model is a one-dimensional continuous elastic string which rests upon a moving frictional surface and has its both ends fixed to rigid supports (see Figure 2-4). Without constructing the model, BK attempt to recount the intuitive unfolding of this system. As an analogy, it is mentioned that this problem is physically equivalent to that of a "stretched violin string bowed by a bow of width almost equal to the length of the string" (343).

Given the difficulties involved in recounting the unfolding of this system with a continuous string, BK end up describing a modified version of the thought experiment. In this

simpler thought experiment, the continuous string is divided into segments of "massless strings attached to point masses" (Figure 2-4). As the string lies upon the moving surface, most of the string rides along the surface without deformation, with the exception of the segments close to the ends of the string. These segments stretch up to a certain point, determined by the elasticity of the segment. After this point, the masses close to the ends slide in the opposite direction of the moving surface. This slide modifies the angle with the next segments of the string and thus stretch them. After a number of slides of the end masses, the second masses are pulled by the end masses. Progressively, more masses get involved in sliding events and, eventually, all masses get involved in large, although infrequent, sliding events (BK: 343). This is a first attempt by BK to describe the dynamics of a continuous system in terms of a discrete system that approximates it.

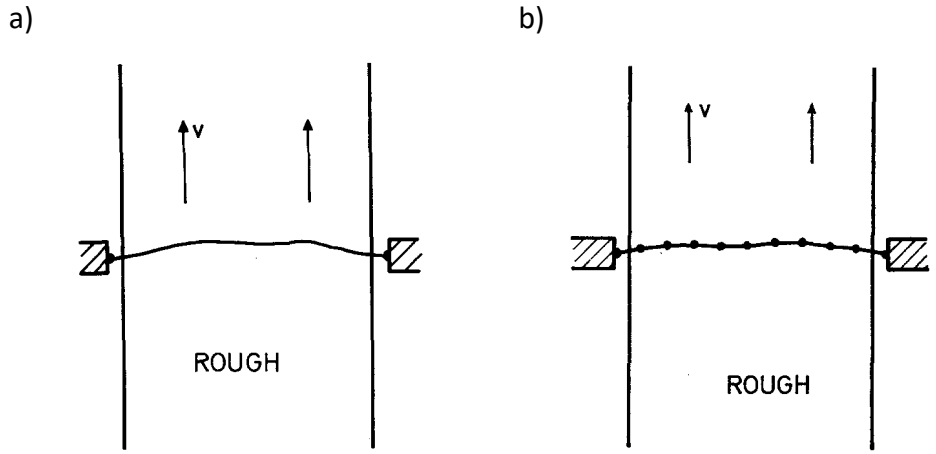


Figure 2-4: Bird's-eye view of BK's thought experiment: a) continuous elastic string; b) discrete elastic string (BK, 1967)

Inspired by this thought experiment, BK begin describing a different but related system, which eventually becomes their laboratory model (or, more precisely, the material implementation of their model). The material implementation of the BK model consists of a linear array of eight identical blocks (each 142 gr), connected through coil springs (see Figure 2-5). Springs with specific elastic coefficients are tested in distinct experiments. BK report two experiments. In one, all the springs have the same elastic coefficients (2×10^5 dynes). In the other, the elastic coefficients are graduated by using coil springs of different lengths. The array of blocks rests on a rough horizontal surface with approximately uniform friction. The

first block's spring is connected through a thread to a motor that pulls the system at a constant low velocity (2 cm/min). The last block is free.

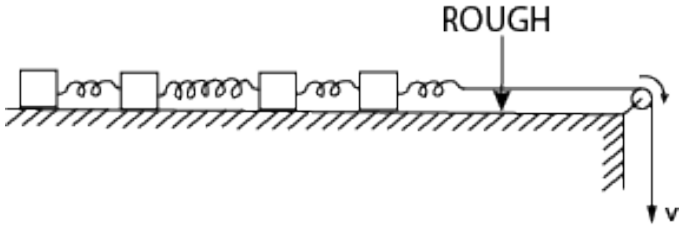


Figure 2-5: Schematic diagram of the BK model in its material (laboratory) implementation. Only four of the eight blocks are depicted (adapted from BK, 1967).

As the motor starts pulling, the first spring starts stretching. Eventually, the tension of the spring exceeds the threshold of the static friction between the first block and the rough surface. Thus, the first block slides reducing the tension on the first spring, but also pulling the second spring, augmenting the tension of the latter. BK refer to the sliding events as "shocks." As the simulation proceeds, eventually all blocks get involved in shocks. The shocks release the elastic potential energy accumulated in springs. Given that the springs are assumed to exhibit Hookean elasticity, the potential energy accumulated in the springs after the mth shock can be characterized as:

$$E = \frac{1}{2}k_1(x_0 - x_{1,m} - l_1)^2 + \sum_{n=2}^{N}\frac{1}{2}k_n(x_{n-1,m} - x_{n,m} - l_n)^2 \qquad (E1)$$

where $x_{n,m}$ is the coordinate of the nth block after the mth shock, l_n is the length of the unstretched nth spring, and k_n is the elastic coefficient of the nth spring. Thus, the potential energy is expressed as a function of the block's coordinates. This way, BK get to know the release of potential elastic energy in each shock, because they can measure the coordinates of the blocks before and after each shock.

BK's laboratory model can be characterized as a preliminary experiment to the mathematical implementation. It plays three practical purposes. First, the laboratory model allows BK to test their physical intuitions derived from their thought experiments. Second, the design of the laboratory model serves as a prototype for the design of the mathematical

implementation. Third, the results of the laboratory model serve as a foil for those of the mathematical model. And, in fact, the results in both implementations are comparable, as the following quote suggests: “The result of the [numerical-computer model] yields a picture of a process which is a generalization of the observations on the laboratory models […]” (359). Now, I proceed to present the mathematical model.

2.2.3 BK Model – Mathematical Implementation

The mathematical implementation has a slightly different, more general setting than the laboratory one. In this case, the modelled system is basically the same as the one in the laboratory model, but all the blocks are additionally connected via flat springs to a moving plate (see Figure 2-6). The moving plate plays the role of the pulling motor in the material experiment. This setting is equivalent to the one in the material model in the case that all flat springs have null elastic coefficient with the exception of the first one.

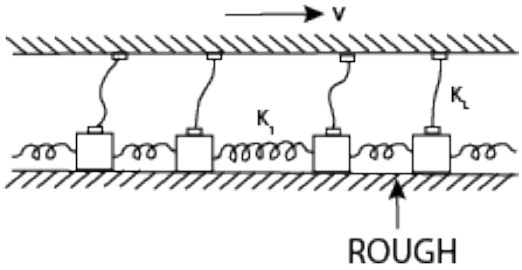


Figure 2-6: Schematic diagram of the BK model in its mathematical implementation. KL: flat spring constant; K1: coil spring constant; V: velocity of the pulling slab (adapted from Ferguson et al. *1998).*

BK engage in the physico-mathematical description of such a generalized version of the spring-block system with N blocks. This theoretical treatment of the model involves: i) characterizing an adequate law of friction and ii) providing the equations of motion of the generalized spring-block system. The law of friction assumes that friction acting upon a block is a function of each block’s velocity. In particular, friction of moving blocks is lower than friction of static blocks.[8] BK include two additional factors into their law of friction to make

[8] BK consider two possible laws of friction. In both, the static friction is greater than the dynamic friction at small velocities. However, in one of them, the static friction is discontinuous with the dynamic friction at small velocities. In the other, the static friction is approached continuously as velocities get closer to zero. In the end, BK choose to simulate a continuous version of the friction law.

their description more "realistic": one due to radiation effects and the other due to viscosity. Roughly, radiation is the energy released during the slide of blocks beyond the boundaries of the spring-block system. And viscosity is a measure of the resistance of a medium to deformation. Frictional forces, radiation, and viscosity are integrated in one physical term in the equations of motion.[9]

The equations of motion are of the form:

$$m_j \ddot{x}_j = T_j - T_{j-1} + T_j^* + F_j^*, \qquad j = 1 \dots N \qquad (E2)$$

where m_j is the mass of the jth particle (i.e., block), T_j is the tension in the spring joining particle j to particle j+1, T_j^* is the restoring force due to the flat spring connecting the jth particle to the moving slab, and F_j^* is a term – function of the velocity of particle jth ($\dot{x}_j$) – that includes frictional forces, radiative effects, and viscosity. More explicitly, the physico-mathematical description of the general spring-block model amounts to the following 2N first order differential equations:

$$m_j \dot{y}_j = u_j(x_{j+1} - x_j) - u_{j-1}(x_j - x_{j-1}) - \lambda_j(x_j - Vt) - E_j y_j + F_j(y_j)$$
$$\dot{x}_j = y_j$$
$$j = 1, 2, \dots, N\ ;\ x_0 = x_1\ and\ x_{N+1} = x_N$$

where $u_j(x_{j+1} - x_j)$ is equal to T_j ; $-\lambda_j(x_j - Vt)$ is equal to T_j^* ; $E_j y_j$ is the force due to radiation of the jth particle ; and $F_j(y_j)$ combines the viscous and frictional forces of the jth

[9] BK's law of friction is defined by parts in the following way:

$$F = \frac{B}{1 - A(v + H)} - Ev\ ; v < -H$$
$$F = -\left(\frac{B}{H} + E\right)v\ ;\ -H < v < H$$
$$F = \frac{-B}{1 + A(v - H)} - Ev\ ; H < v$$

where A, B, E and H are constants. A stands for the effects of viscosity. B stands for the static friction. E stands for the effects of radiation. H is a critical value in velocity above which the effects of viscosity are manifested. Negative values for *v* mean velocities in the negative direction.

particle. x_j is the coordinate of the jth block; μ_j is the elastic coefficient of the j[th] coil spring; $-\lambda_j$ is the elastic coefficient of the j[th] flat spring; V is the velocity of the moving slab.[10]

BK do not solve these equations analytically. Instead, they solve them numerically, using a Runge-Kutta method, deployed in a computer program. (The computer program was written by collaborator Susan Karman.) This way, the potential energy is obtained in different iterations which stand for moments in the evolution of the system. This endeavor instantiates what BK refer to as "model seismicity": they develop an analog computer model that solves the theoretical problem. The motivations for engaging in this method are practical: "The problem of solving a number of simultaneous ordinary non-linear differential equations is best left to an electronic computer" (BK: 351).

BK's mathematical model has advantages over their laboratory model. In particular, by analyzing the spring-block system mathematically, BK can introduce more variables and gain manipulatory control over them. This is expressed in the following quote: "We describe […] the formulation of the mathematical problem, suitable for computation, which is a generalization of the laboratory model. By leaving the laboratory model at this stage and proceeding to a mathematical analysis of the model we are able to introduce and vary parameters governing the features of the model" (BK: 351). Beyond gaining more control over the simulations, BK can also simulate features of earthquakes that were not suitable to study in the laboratory model (see discussion on "aftershocks" below). Having described the laboratory and mathematical implementations of the BK model, I now proceed to present the main results of the simulations.

2.2.4 Main Results of the BK Model

BK report results of two simulations in the laboratory implementation of their model. One simulation is run with springs with equal elastic coefficients. The other simulation is run with springs with graduated elastic coefficients. In the latter case, the spring with the smallest coefficient is located closer to the motor and the one with the largest coefficient is connected

[10] Note that the constants in the law of friction (i.e., A, B, E, and H) are specific for each block (i.e., A_j, B_j, E_j and H_j for j from 1 to N).

to the free block. The simulations start with the springs approximately unstretched. The motor pulls the system at a constant rate of 2 cm/min. Several hundreds of shocks of different sizes are registered in each simulation in the curse of an hour.

I proceed to present and discuss qualitative features of the simulation results. I restrict this presentation and discussion to the results of the simulation with springs with equal coefficients. This is permissible, given that the qualitative features of both simulations are alike: "The results [of the simulation with unequal springs] are quite similar to those for the case of the equal springs" (BK: 349). To begin with, BK report a "charging cycle" for the spring-block system, which is a period of relatively small shocks that load potential energy into the system. BK plot the charging of the system in a "potential energy" versus "time" diagram (see Figure 2-7). Accumulated potential energy in the spring block system is calculated with equation *E1* above. Given that the rate of pulling is constant, time is expressed in terms of the horizontal coordinate of the starting point of the system, i.e., the junction between the first spring and the motor. The simulation shows that progressively larger shocks occur. After a certain point, shocks involving the eight blocks are periodically observed. BK describe such periodicity as an "oscillation" between an upper and lower threshold for potential energy in the system, driven by the pulling of the motor.

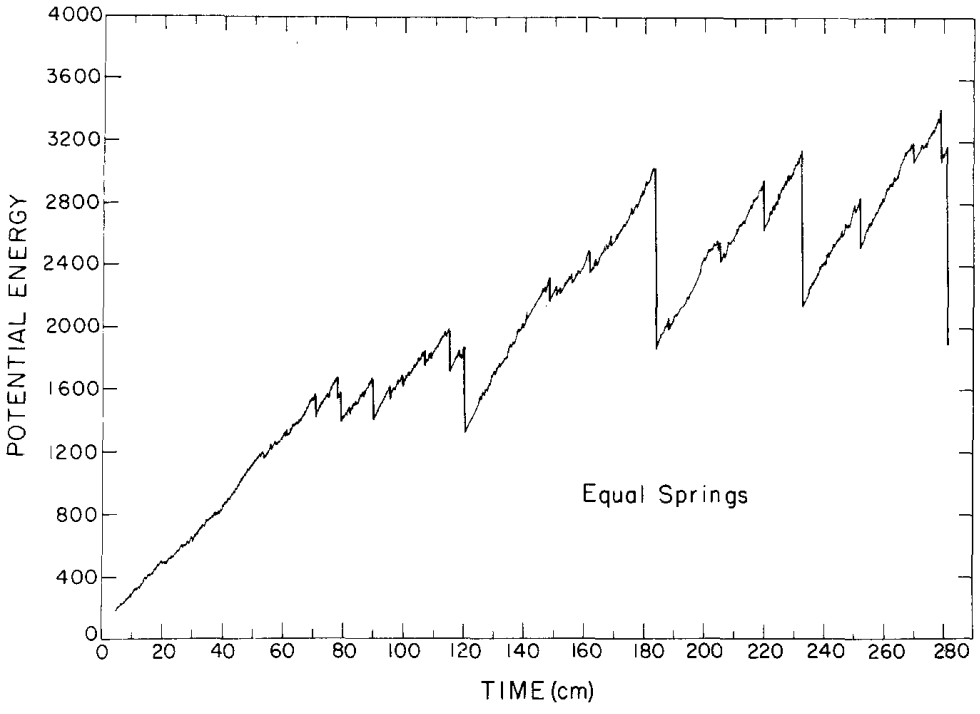


Figure 2-7: Accumulated potential energy as a function of time. Time is expressed in length units because of the constant rate of pulling by the motor (BK, 1967).

BK also report the relation between size of shocks and their frequency. The size of shocks is expressed in terms of the release of potential energy (represented as the length of vertical segments in Figure 2-7). By plotting the results in a log-log diagram, BK report a slope of -1 over a wide range of energy releases (see Figure 2-8). Such slope reflects a power-law relation between size and frequency of shocks. BK compare this result to the behavior of real earthquakes and note the resemblance with the GR law.

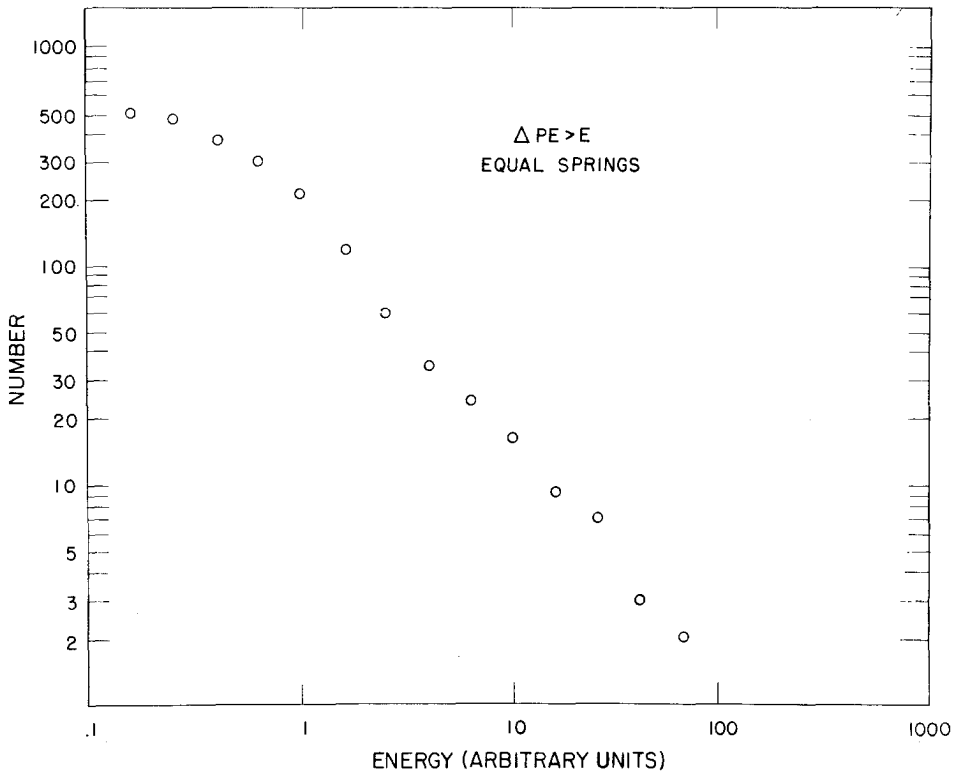


Figure 2-8: Frequency-energy diagram (BK, 1967).

Despite obtaining power-law distributions in the simulation, BK emphasize that there are dissimilarities between the exponents in the simulated power-laws and those reported in real sets of earthquakes. Consider the following quote:

> *"For many regions of the earth, b is approximately 1 in large shocks. However, our model probably better simulates small shocks than large. For shocks with M < 7.1, b = 0.58 (Richter, 1958). The coefficient β is poorly fixed; we shall use the value suggested by Richter (1958) and Båth (1958) of β^1 = 1.5. The product bβ for small shocks is about 0.4, a*

value approximately one-half the laboratory results. There are uncertainties in both the numbers b and β for naturally occurring events; for example, one may note that the earlier values of β reported in the literature were as high as 2.0. This is inadequate to account for the difference in the two values. A more tantalizing possibility is that this difference reflects the differing dimensionality of the fault surface for the laboratory and the field cases; this possibility is a matter of conjecture at present" (BK: 347-8).

Besides dissimilarities in terms of the exponents, deviations from power-law behavior are reported at the high and low energy portions of the spectrum. Although this feature is also reported in the distributions of naturally occurring earthquakes, BK intend to offer explanations of this feature that are based on characteristics that are idiosyncratic to the simulation. Consider the following quote: "Because of the finite number of masses there are no quakes smaller than a certain critical size. In addition, we may miss recording a number of the smallest quakes. Similarly, because of the finite number of masses involved in the interaction, no shocks which release more than a certain energy will ever take place" (BK: 348).

BK also study the degree of randomness of the occurrence of shocks in the laboratory model. BK claim that the simulation results fit a Poisson distribution with a confidence level below 10%. This means that shocks are not random, causally independent events. Rather, shocks are causally relevant to other shocks. This should come as no surprise. A shock affects the stretching of various springs, modifying the initial conditions for next events.

With regards to simulations with the mathematical model, BK report the numerical results of a simulation for a system with ten blocks. The ten blocks are divided in three sections, each one with specific values for the parameters of their respective friction law (Figure 2-9). BK describe these three sections with an explicit representational intention: in one end, five blocks form an "analog" of a strong seismic fault, with low values for viscosity (P-fault). The middle section, with two blocks, forms a viscous region that transmits the effects of the two extreme sections. In the other end, three blocks form a second fault, similar to the large one but with lower values for static friction (A-fault).

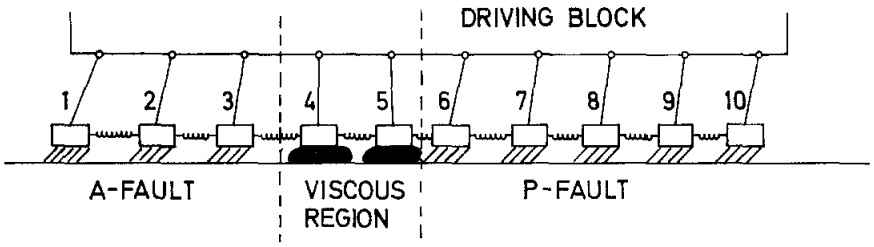


Figure 2-9: Modeled system for numerical simulations (BK, 1967).

BK claim that the numerical results yield "a picture of a process which is a generalization of the observation on the laboratory model described earlier" (BK: 359). Because of this, I do not go into much detail discussing these results. However, one important distinction is worth mentioning: the presence of "aftershocks." Basically, aftershocks are a series of minor to medium shocks that follow a large shock and are causally related to it. BK are able to simulate aftershocks in the mathematical implementation because in this case they can economically introduce variables that stand for viscosity in a geological fault. As an illustration, BK report a major shock in the P-fault section. The duration of such shock is of 1.2 arbitrary unit of time. The shock affects the viscous section, which transmits the effects of the main shock to the A-fault section. After $5x10^4$ units of time, a shock is reported in this second section. The time delay is significantly large to be deemed part of the same shock. However, the second shock is causally related to the main one and it is thus considered an aftershock.

These are the main results obtained by BK. I am particularly interested in the power-law relation between frequency and size of earthquakes for two reasons. First, this is a pattern that is also simulated in the OFC model. This allows for direct comparisons between the BK and OFC models. Second, power-law patterns have been an object of study by a research program in physics known as "self-organized criticality" (SOC). The OFC model explicitly responds to this program. It is worth mentioning that BK simulated the power-law behavior of earthquakes two decades before the SOC program came to prominence. Now, I proceed to examine the OFC model.

2.3 Case Study II: the Olami- Feder-Christensen (OFC) Model

Olami, Feder, and Christensen's 1992 paper – "Self-Organized Criticality in a Continuous Nonconservative Cellular Automaton Modeling Earthquakes" – is a further contribution to the study of the occurrence of earthquakes. In this paper, OFC introduce a new model that simulates features of the behavior of earthquakes. Most notably, the OFC model simulates the power-law distribution of earthquakes with regards to their size. This amounts to saying that the OFC model predicts the GR law. In addition to OFC's 1992 original paper, Christensen and Olami published two more detailed papers that same year: a) "Scaling, phase transitions, and nonuniversality in a self-organized critical cellular-automaton model"; and b) "Variation of the Gutenberg-Richter b values and nontrivial temporal correlations in a spring-block model for earthquakes." In these papers, Christensen and Olami discuss further details concerning the OFC model, especially the variance in b-values (i.e., exponents in the power-law) as a function of the conservation regimes in the model. Hence, I use Christensen and Olami (1992a) and Christensen and Olami (1992b) as natural extensions of Olami, Feder, and Christensen (1992).

Within the first page of their original paper, OFC acknowledge two main scientific endeavors which influence their own research. The first one is a research program known as "self-organized criticality" (SOC).[11] This program was originally developed in statistical mechanics and condensed matter physics, but it soon found affinities with other disciplines. One of these disciplines is seismology. Bak and Tang (1989) submit an explicit and strong affinity between earthquakes and SOC: the thesis is that earthquakes *are* a self-organized critical phenomenon. OFC submit a slightly more moderate thesis: "[t]he dynamics of earthquake faults *may* provide a physical realization of the recently proposed idea of [SOC]" (OFC: 1244; my emphasis). OFC add that earthquakes are "probably the most relevant paradigm of self-organized criticality" (OFC: 1244).

In addition to the SOC program, the OFC model is highly influenced by the BK model. As OFC assert, the OFC model "is directly mapped into a two-dimensional version of the famous Burridge-Knopoff spring-block model for earthquakes" (1244). OFC also claim that

[11] Frigg (2003) characterizes SOC as a research program in Lakatos's sense. In particular, Frigg claims that SOC has a hard core conformed by a set of "paradigm models" (626).

their model is *equivalent* to a quasi-static two-dimensional version of the BK model (1244). In this sense, the OFC model can be characterized as a model of the BK model: the OFC model extends the features of the BK model and implements them in a distinctive way, namely as a cellular automaton (see discussion below). To be sure, OFC are not the first to deliver a cellular automaton model of earthquakes based on the BK model (e.g., see Otsuka, 1972). Their peculiarity, however, is that they do so within the context of the SOC program. Furthermore, the OFC model is arguably the most comprehensive and succinct version of the BK model explicitly designed for the SOC program as a cellular automaton (cf. Pruessner, 2012: 126). Given its centrality to OFC's research, I begin my discussion of the OFC model with three preliminary sections on fractals and power-laws, the main tenets of the SOC program, and SOC models with cellular automata.

2.3.1 Fractals and Power-Laws

Originally, SOC was introduced as "the common underlying mechanism" responsible for various instances of spatial fractals and power-law temporal and spatial correlations in dynamical systems (Bak *et al.* 1987: 381). In this sense, the motivations of the SOC program can be characterized as the search for a theory that accounts for the physics of fractals (cf. Bak & Chen, 1989). Physical fractals exhibit a remarkable attribute, namely (statistical) self-similarity or scale invariance. Self-similarity in physical fractals can obtain in space or time, although some authors restrict the term 'fractal' to refer to spatial phenomena. Temporal self-similarity is often referred to in terms of "1/f noise" and other related concepts.

Mathematical treatment of physical fractals was pioneered by Benoit Mandelbrot. He realized that scale invariant systems can be mathematically represented as power-laws. Consider a pedagogical example introduced by Mandelbrot in his 1967 paper "How long is the coast of Britain? Statistical self-similarity and fractional dimension." The motivation is the following: Seacoasts – among other geomorphological features – look similar at different scales of observation. In other words, they are (approximately or statistically) self-similar. Because of this feature, it is not straightforward to measure the length of a seacoast: if we measure the coast with high precision, the measured length is larger than the one measured with less precision. Accordingly, length is not an expedient feature to characterize self-similar

curves. Mandelbrot suggests that a better, more faithful attribute is the exponent of such curves when expressed as a power-law.

To unpack this, consider the following thought experiment. Suppose that you want to measure the coast of Britain with a 200km ruler. You go around the coast of Britain taking measurements, with the proviso that in each measurement both ends of the ruler must touch the coast. If you cut the ruler in half and do the measurement again, with the same proviso, you find that the measured length of the coast is larger (i.e., you use the ruler more times). If you keep cutting the ruler in halves and do new measurements, each new length of the coast is larger than the previous one (see Figure 2-10). The relation between the length of the ruler (r) and the number of times you use the ruler (N), follows the power-law:

$$N = Cr^{-D}$$

where C is a constant and D is the exponent of the power law, also referred to as the fractal dimension of the system. It is in this sense that self-similar objects can be described by power-laws, and the exponents of the power-laws are a reliable attribute of such objects. For self-similar systems, D is a fractional number: that is the origin of the term 'fractal'.

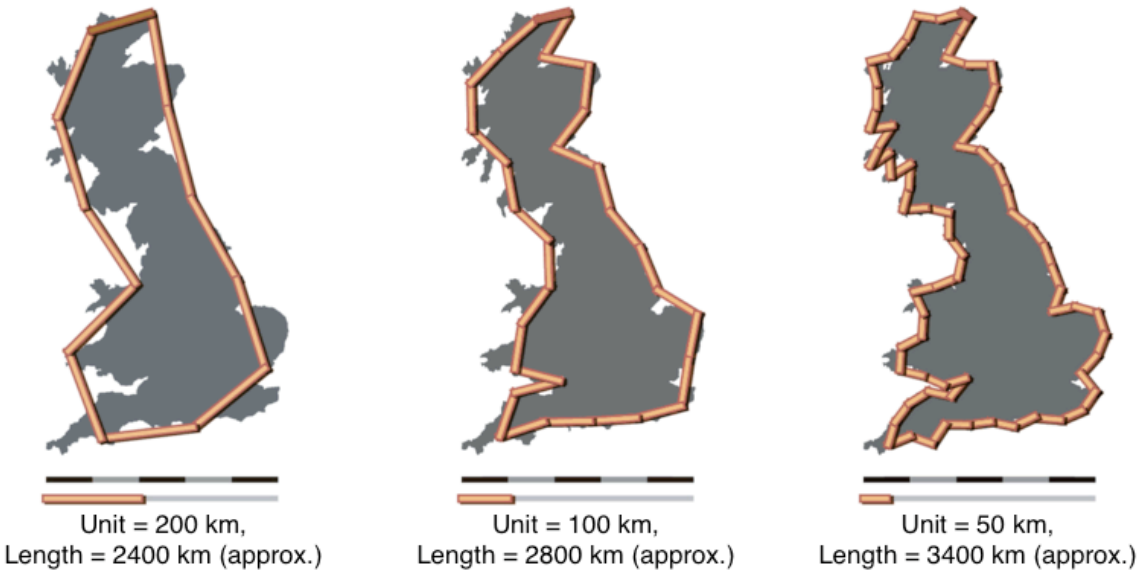


Figure 2-10: Length of the coast of Britain, measured with a ruler progressively smaller by half. Note how the number of times the ruler is used increases inversely with the size of the ruler (Source: Wikimedia Commons).

Scale invariance is not only restricted to the self-similarity of objects at different scales of observation (i.e., resolution). In addition, scale invariance can also be embodied in the distribution of sizes of a set of objects. A set of objects exhibits fractal or scale-invariant size statistics if the cumulative distribution N(r) of the objects sizes is a power law distribution:

$$N = Cr^{-D}$$

where C is a constant and D is the fractal dimension of the distribution. The absence of a characteristic scale in power-law distributions makes them scale invariant (Hergarten, 2002: 14-5). It is in this last sense that fractals and scale-invariance are relevant concepts to characterize the behavior of earthquakes: earthquakes sizes follow a power-law distribution of the form:

$$N = N_0 E^{-B}$$

where *N* is the number of earthquakes above size *E*, and *B* is equivalent to $b\beta$ (see above). Having clarified the notion of "fractals" and their relation to power-laws, I proceed to deliver a first approach to the main tenets of the SOC program.

2.3.2 The Research Program of Self-Organized Criticality

The origins of self-organized criticality, as a research program, are consistently attributed to two seminal papers by Bak, Tang and Wiesenfeld (hereinafter BTW): 1987's "Self-Organized Criticality: An Explanation of 1/f noise" and 1988's "Self-organized criticality."[12] These papers have been extremely stimulating. Hergarten (2002) reports that, in the 1987-1997 decade, more than two thousand articles on SOC were published, making BTW (1987) the most cited paper in physics during this period (ibid: 88). Watkins *et al.* (2016) also report that BTW have received more than 6600 citations between 1987 and 2016 (4).

[12] Credit to BTW is well justified: the concept of 'self-organized criticality' is introduced in BTW's paper. However, previous scientific research had dealt with simulations of phenomena that would eventually be recognized as examples of SOC. A notable example is BK's paper itself, which – I suggest – can be considered as a precursor to the SOC program. Indeed, BK achieve something that the SOC program aims for, namely the simulation of power-law behavior of complex system via models of discrete interacting entities. Furthermore, BK show that the power-law behavior of earthquakes is robust with regards to (some sorts of) manipulations performed upon the simulation's setting.

Researchers in this program are concerned with the dynamics of natural systems – henceforth SOC systems – that share a common set of features. First, SOC systems are composed of several interacting parts which typically exchange some physical quantity, like force or energy. Second, the dynamics of SOC systems is shaped by interactions with their environment.[13] That is, SOC is concerned with dissipative systems: open systems that exchange energy, matter, or force with their surroundings. These two first features can be summarized by saying that the SOC program studies the evolution of systems under the influence of two main factors: external driving and internal interactions. Third, SOC systems are governed by local thresholds. That is, there are local limits for the amount of matter, force, or energy that an entity, a set of entities, or a spatial domain can accumulate. Whenever the thresholds are exceeded, the surpluses are redistributed to the immediate surroundings, i.e., neighbor entities. As these exchanges unfold, neighbor entities might exceed their own thresholds, thus producing chain reactions of varying size, also known as "avalanches." Fourth, the rate in which matter, energy, or force are introduced or exerted upon the system from the environment is orders of magnitude lower than the rate in which internal interactions (avalanches) unfold. This is usually referred to as the "separation of timescales" condition for SOC systems.

The basic tenet of the SOC program is that dynamical systems which exhibit the aforementioned features self-organize towards, and fluctuate about, a quasi-equilibrium state. In this state, interactions between the component parts of the system can trigger chain reactions that involve a variable amount of parts. In other words, SOC behavior involves avalanches of different sizes. A robust feature across SOC systems is that the distribution of avalanche size follows a power-law. This feature is analogous to the power-law behavior of thermodynamical systems at a critical point near a phase transition. Hence the name self-organized criticality.

[13] This claim presupposes that the distinction between what is internal and external to the system is unproblematic. However, this is not straightforward. Deciding the boundaries of a system is a difficult task. Modeling often plays a major role in refining the conception of SOC systems and deciding where their boundaries lie.

The original motivations of the SOC program have evolved with the passing of time, especially due to new disciplines entering the field. From a modest beginning in statistical mechanics and condensed matter physics, the SOC program soon captured the interest of scientists in dissimilar fields. For example, the SOC program has reached disciplines such as seismology, astrophysics, ecology, neuroscience, and sociology, to name a few (Watkins *et al.* 2016: 4). As a consequence, the commitments of the SOC program have become heterogenous and often unclear. To better deal with this scenario, Watkins *et al.* (2016) deliver a revisionary paper concerning the SOC program.[14] They analyze the main concepts in the program and attempt to settle the main controversies. I follow their account of the state of the art in SOC research for the rest of this overview and complement it with additional sources when needed.

Watkins *et al.* (2016) submit that BTW's original motivations amount to providing a dynamical theory of the physics of fractals. SOC is thus introduced as a mechanism responsible for fractals. However, there seems to be a mismatch between what BTW actually did in their papers and the breadth of their conclusions. BTW study the dynamics of power-law *correlations* in spatial fluctuations, manifested as chain reactions or avalanches. This is different from studying the power-law distribution of *sizes* in nature, i.e., fractal systems.[15] Power-law correlations may cause fractal patterns (if other conditions obtain). However, it is not the case that all fractals are produced by power-law avalanching. This is the core of the problem: BTW use power-law distributions of sizes (fractals) as a proxy for power-law correlations (avalanches), but such proxy is fallible. Watkins *et al.* (2016) express this problem clearly:

[14] A special edition of the Space Science Reviews (Volume 198, Issue 1-4, January 2016) commemorates the 25-year anniversary of self-organized criticality. Watkins *et al.* address general issues. Additional articles address more technical topics and applications of SOC to astrophysics.

[15] Hergarten (2002) reflects upon a related issue, namely the different notions of scale that exist. In a section of his book entitled "Fractals or Fractal Distributions," Hergarten explicates that there are two different definitions of scale invariance: "The first definition is based on the geometric properties of a set in a n-dimensional Euclidian space, while the second definition focuses on sizes of objects and disregards any further properties such as spatial alignment. On a less formal level, the difference between both definitions can be attributed to different *notions of scale.* When the distribution of object sizes is considered, we focus on a scale defined by the object we are actually looking at. In contrast, the scale in the geometric definition is some kind of resolution […]" (17). I suggest that, in light of Watkins et al. and Hergarten's comments, there seem to be at least three kinds of power-laws in consideration: i) the power-law of geometrical properties; ii) the power-law of distribution of sizes; and iii) the power-law of correlations (or avalanches). These power-law regimes can coincide in some contexts, but that is not necessary.

> *"The gap between the relatively specific idea of explaining space-time fractal avalanching phenomena and therefore spatio-temporal correlations, and the aspiration that many perceived to explain any fractal in space or time, or even any power law distribution, has been a perennial problem, and a key source of the controversy and misunderstandings that still surround SOC" (11).*

Watkins *et al.* (2016) submit that there are four key claims in the early formulations of the SOC program, as intended by BTW. These claims were shaped by BTW's explicit motivation of explaining spacetime fractals by means of SOC as a hypothetical mechanism. Some of these commitments have been abandoned, others have been refined:

1. Spatial and temporal scaling must be connected. This claim, however, has been proven false (ibid: 12). Temporal fractality can cause spatial fractality, and vice versa. But there is no necessity in their relation.

2. Dissipative systems organize themselves towards a critical state in which power-law correlations occur. This claim is controversial: it uses the term 'critical' in a new way. Traditionally, criticality in continuous phase transitions obtains in a singular point in the parameter space. This point is reached by tuning a control parameter. As a familiar example of control parameter, consider the temperature in the classic Ising model. In contrast, the critical state in SOC is not (finely) tuned: systems organize themselves towards the critical state.[16] Also, the critical state in SOC is not a singular point in which power-law correlations occur. Rather, the power-law correlations occur as a consequence of the system fluctuating about this state. In this sense, the critical state in SOC is better characterized as an attractor. Another aspect of SOC that recalls the criticality of systems in continuous phase transitions is their divergent susceptibility. This basically means that the size of avalanches spreading across a system diverges with the size of the system and any perturbation, even small ones, can trigger system-size avalanches. BTW thus use the term 'criticality' in a strictly

[16] There has been debate concerning the issue of tuning. The problem can be spelled out in the following way. There is consensus among SOC researchers upon a minimal set of conditions that need to be obtained in order for a system to exhibit SOC behavior. In this sense, there is a minimal tuning concerning the features of SOC systems: one needs these conditions for SOC to obtain. However, there is no need to tune variables that resemble control parameters in other thermodynamical systems. As a consequence, it is often said that SOC lacks the need of "fine" tuning.

phenomenological way: i) SOC systems display power-law correlations that *resemble* those of thermodynamical systems at a critical point; ii) SOC systems exhibit divergent susceptibility to external perturbations, like thermodynamical systems in the critical point.

3. In order to exhibit SOC behavior, a system must be open, extended, dissipative and slowly driven. This claim has been the focus of much research and it has been continuously refined. A later formulation by Jensen (1998) has become an agreed standard: the dynamic of systems exhibiting SOC is slowly driven, interaction dominated, and with local thresholds (SDIDT).

4. Spacetime fractals are snapshots of SOC. This claim is misleading: it has been shown that SOC is not the only mechanism responsible for all spacetime fractals in nature (see e.g., Sornette, 2006; Jensen, 1998). However, *some* spacetime fractals are the consequence of SOC. In spite of its misleading nature, this assumption played a crucial role in configuring the original commitments of the SOC program.

After reviewing the key claims of the original SOC program, it seems accurate to state that the core of the SOC program has become the study of the conditions under which self-organized critical behavior occurs in complex systems and the resulting behavior. In other words, there are two main concerns in SOC research: causes and effects. Confusingly, the term SOC is often used to refer to one or the other. On the one hand, SOC is a mechanism – or set of causal factors – responsible for a robust and general behavior. On the other hand, SOC is used to refer to the resulting behavior itself. This is admittedly unclear: the term SOC is used to refer to an explanandum (SOC behavior) and explanans (SOC mechanism). To overcome this ambiguity, Watkins *et al.* (2016) make a distinction between the phenotype (i.e., features of SOC behavior) and genotype (i.e., SOC as a set of causes).

SOC behavior (i.e., the phenotype) amounts to three key features. First, SOC behavior includes non-trivial scaling. This basically means that systems engaged in SOC behavior exhibit self-similarity or scale invariance. Second, SOC behavior involves spatio-temporal power-law correlations. That is, avalanches' size follows a power-law regime. Third, SOC behavior comprises the self-tuning of the system in question to the critical state. Hence the name self-organized criticality.

There are three key causes of SOC behavior (i.e., the genotype) which must occur together for SOC behavior to obtain. First, in order to exhibit SOC behavior, systems must engage in non-linear internal interactions. In the case of SOC, this amounts to the existence of local thresholds for physical quantities. Once these thresholds are exceeded, exchange with the immediate surroundings occur. This leads to the second key cause of SOC: avalanching. Avalanches are chain reactions in which exchange of matter, energy, or force occurs among constituent parts of a system.[17] Third, SOC requires a separation of timescales between the external driving and the internal interactions. More explicitly, external driving is slow and internal exchanges (or "relaxation") is fast, almost instantaneous from the perspective of the driving.

Although there is consensus with regards to the crucial role of these three causal factors, they do not exhaust the causes of SOC, as other studies suggest. Two additional factors are of interest to my argument and I discuss them below. First, the degree of conservation of the physical quantity exchanged in internal interactions affects the obtaining of SOC. The importance of this factor is actually reported in OFC's study. They show that there is a limit for the amount of dissipation of energy in systems that exhibit SOC behavior. Second, the topological features of the system in question might affect the obtaining of SOC. This means that how the entities within the system are organized in their interactions is causally relevant. Researchers have shown that some topological configurations, *ceteris paribus*, do not trigger SOC behavior. For example, Caruso *et al.* (2006) explore the role of different topologies in the context of the OFC model. They report that, *ceteris paribus*, scale-free topologies do not trigger SOC behavior. Having presented an overview of the SOC program, I proceed now to discuss one of its most consistent methodologies, namely computer simulations with cellular automata.

2.3.3 SOC Models and Cellular Automata

[17] Watkins *et al.* (2016) reflect upon the double status of avalanches as cause and effect. On the one hand, avalanches are required for the system to self-organize towards the critical state. In this sense, avalanches are a cause of SOC. On the other hand, avalanches are also a symptom of a system being engaged in SOC. In particular, this is expressed in the second phenotypic feature of SOC introduced above, i.e., spatio-temporal power-law correlations.

Modeling plays a crucial role in the SOC program. Most SOC models are computer simulations designed as cellular automata. OFC acknowledge the role of cellular automata in the SOC program. Consider the following quotes: “The idea of SOC went hand in hand with a definite cellular-automaton algorithm, suggested by BTW, initiating enormous scientific activity. Various different cellular automata were studied, and a great effort was spent on deriving some theoretical understanding of the new models” (Christensen & Olami, 1992a: 1829). And: “[t]he study of the SOC systems has to a great extent been based on simulations using cellular automaton models” (OFC: 1244).

In order to discuss cellular automata, I resort to an encyclopedia entry by Berto and Tagliabue (2017). They highlight three key features of cellular automata, namely i) their discreteness; ii) abstractness; and iii) computational capacities. First, cellular automata are (typically) discrete in the sense of being composed of individual entities, referred to as cells. At any moment in a simulation, cells instantiate a state, which amounts to a (numerical) value. In this sense, cells can be characterized as variables that take distinct values as the simulation progresses. Second, cellular automata are abstract in the sense of being characterized by their structure and instructions in purely mathematical terms. Often, cellular automata are implemented in physical systems, especially in computer programs for purposes of computer simulation. However, a particular implementation is not part of the identity of a cellular automaton. Third, cellular automata are computational systems: they are capable of computing functions and solving algorithms.

To illustrate the basics of cellular automata models applied to SOC research, Figure 2-11 displays four steps of a simulation involving avalanches in a simple 5 x 5 lattice cellular automaton. Each cell has attached a numerical value, which changes in successive steps according to update rules. The update rules can be divided into external drive, or simply “driving,” and internal dissipation, or simply “relaxation.” The driving rule corresponds to an increment of one unit (+1) in a random cell. The relaxation rule corresponds to the following: if a cell has a number greater than 3, then subtract 4 to the cell and add 1 to the perpendicular neighboring cells. In step 1, the driving rule applies an increment (+1) at the central cell. Then, the relaxation update rule operates in consecutive steps, until quiescence is reached. In general, relaxation update rules can be thought of as instructions to distribute a quantity

among neighboring cells when a threshold is reached. Relaxation update rules are recursive, i.e., they are repeatedly applied until a stable configuration or quiescence is reached – in this case, until all cells have a number equal or less than 3. These consecutive updates by relaxation stand for the so-called "avalanches." Avalanches are characterized by their duration and size. Duration is the number of steps it takes the cellular automaton to reach a new stable configuration. Size is the number of cells involved in update. In this example, the depicted avalanche has a duration t = 3 (quiescence is reached after three steps), and a size s = 13 (thirteen cells are updated). Importantly, as a matter of design, while an avalanche is unfolding, there are no further increments. This feature amounts to the so-called "separation of timescales." In other words, the external drive proceeds at the macroscopic timescale, very slowly, while relaxation – and thus avalanches – occur at the microscopic timescale, i.e., instantaneously from the perspective of the macroscopic timescale.

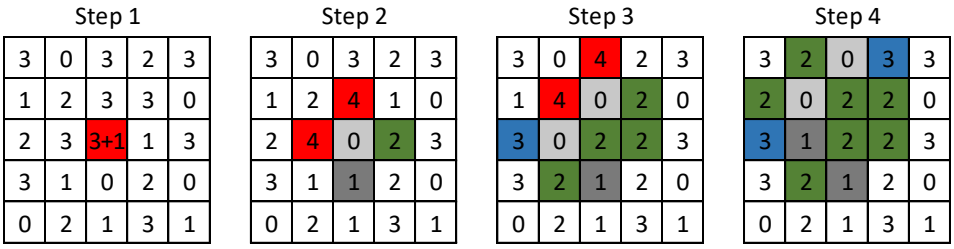


Figure 2-11: Four steps in the evolution of a 5x5 lattice cellular automaton (adapted from Gros, 2015: 197).

Cellular automata are often used for "phenomenological" modeling (cf. Barberousse *et al.* 2007). This means that cellular automata are used to represent their target phenomena "directly," i.e., without the mediation of an explanatory theory about the phenomena. In practice, this amounts to designing rules in the cellular automaton that correspond to actions or events in the target system. These rules are localized, i.e., directed to the cells, addressing their evolving states and interactions. In other words, rules are intended to capture the lower-level or microdynamics of the target phenomenon. The design of these rules involves assumptions – often highly idealizing ones – concerning the microdynamics of the target. Because of this, correspondence between designed rules in the cellular automaton and local events in the target is subjected to pragmatic standards of minimal resemblance.[18] However,

[18] Barberousse *et al.* (2007) comment on how the representational capacities of cellular automata must be carefully managed: "[Cellular automata's] striking mathematical and logical properties, as well as the powerful visualizations they allow for, should not hide the fact that in order to have them represent the evolution of

it is also the case that scientists may use cellular automata to explore scenarios that do not resemble the actual target. Furthermore, scientists often investigate cellular automata and their features as objects of study in themselves. In this sense, cellular automata can be used in a at least these two fashions, namely representational and non-representational.

The local rules in cellular automata typically give way to complex macroscopic behavior. These macroscopic behaviors are not an explicit part of the design of the program. In technical terms, the complex macroscopic behavior "emerges" from the microdynamics, which is described in terms of explicit local rules. In the words of Wolfram (2002): "very simple rules produce highly complex behavior" (ibid: 39). It is in this sense that cellular automata do not rely on a theory at the level of the macroscopic behavior that they simulate. Rather, they rely on assumptions at the level of local instructions.

In the very title of their paper, OFC characterize their model as a continuous cellular automaton. However, more purist scientists are not content with the employment of the term 'cellular automaton' to describe the OFC model. Strictly speaking, cellular automata are discrete space-time lattices, with discrete state variables. While the OFC model is discrete in space, it lacks uniform discrete time steps and the state variables are continuous. Grassberger (1994) suggests that the right term to characterize a model like the OFC model is 'coupled map lattice' (2436). Nonetheless, the term 'continuous cellular automaton' well caught on (cf. Pruessner, 2012: 127). I use the term 'cellular automaton' in this wider sense, mimicking OFC's attitude. After having addressed the main features of cellular automata in SOC models, I present the OFC model and its main features.

2.3.4 Main Features of the OFC Model

The OFC model is inspired by the BK model (OFC: 1244). More explicitly, the OFC model is a sort of "spring-block" model used as a surrogate system to study the dynamics of earthquakes. OFC design their model as a two-dimensional version of the BK model.[19] In OFC's

natural or social systems, their particularities have to be carefully tamed and the various representational relationships they are involved in have to be no less carefully established."

[19] Otsuka (1972) extended the BK model to a two-dimensional version two decades before OFC.

version, each block is connected to four orthogonally neighboring blocks via coil springs.[20] The array lies upon a rough static plate. All the blocks are connected via flat springs to a moving plate with constant low velocity. The relative motion of the plates forces the various springs to stretch or compress (Figure 2-12).

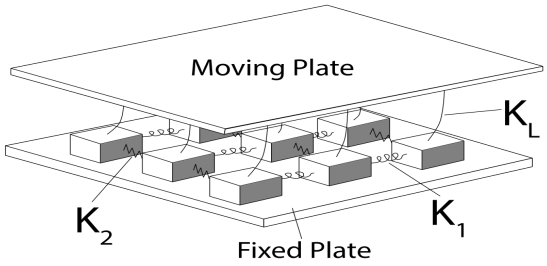


Figure 2-12: Spring-block system assumed for the OFC model, with emphasis on the moving plate, fixed plate, and elastic coefficients along the x-axis, y-axis and z-axis (K_1, K_2 and K_L, respectively). Adapted from OFC, 1992.

In the following, I describe the dynamics of this two-dimensional spring-block system. An *L x L* array of blocks is assumed. Each block has coordinates (*i,j*) – *i* and *j* are integers between 1 and *L* (see Figure 2-13).[21] The displacement of each block from its relaxed position is defined as $dx_{i,j}$. The tensional forces acting upon block (*i,j)* – $F_{i,j}$ – are described as it follows (see Figure 2-13 for reference):

$$F_{i,j} = K_1\left[2dx_{i,j} - dx_{i-1,j} - dx_{i+1,j}\right] + K_2\left[2dx_{i,j} - dx_{i,j-1} - dx_{i,j+1}\right] + K_L dx_{i,j}$$

where K_1 is the elastic coefficient of springs in the x-axis, K_2 is the elastic coefficient of springs in the y-axis, and K_L is the elastic coefficient of springs in the z-axis (i.e., the flat springs). Thus, the OFC model can be used to analyze anisotropic cases, i.e., cases in which the elastic deformation in one direction is different from the one in other directions. In the end, OFC report simulation results assuming isotropic systems, i.e., K_1 is equal to K_2 (but see Christensen and Olami, 1992a).

[20] Depending on the nature of the boundary conditions, the coordination number (i.e., number of neighbors) of blocks on boundaries (sides and corners) can vary.
[21] The OFC model can, in principle, operate in non-equilateral arrays. In particular, an L x 1 array corresponds to the one-dimensional setting of the original BK model.

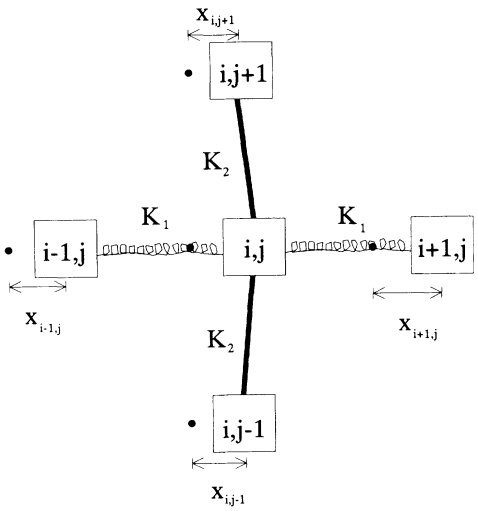


Figure 2-13: Bird's-eye view of a block neighborhood. The view is centered on block (i,j) and depicts its respective neighbors and the springs connected to it with their respective elastic coefficients (OFC, 1992).

As the moving plate pulls the flat springs, the tensional force on each block increases. Eventually, the tensional forces acting on a block exceed the static friction – referred to as F_{th} – and the block slips. As a consequence, the tensional forces in the neighborhood of the slipping block reconfigure. All the tension in the slipping block is redistributed among its neighbors. The amount of force distributed to the neighbors of a slipping block *(i,j)* is a portion of the total force on the slipping block. Such portion is defined as:

$$\delta F_{i\pm1,j} = \frac{K_1}{2K_1 + 2K_2 + K_L} F_{i,j} := \alpha_1 F_{i,j}$$

$$\delta F_{i,j\pm1} = \frac{K_2}{2K_1 + 2K_2 + K_L} F_{i,j} := \alpha_2 F_{i,j}$$

where α_1 and α_2 are the elastic ratios in the x-axis and y-axis respectively. For the isotropic case in which α_1 is equal to α_2, OFC use simply α as a global elastic parameter. Any residual force not distributed to the neighbors is dissipated. As the force acting upon the neighboring blocks increases, it can also reach the threshold for static friction, causing them to slip as well. Thus, this process has the potential to produce chain reactions involving a varying number of slipping blocks.

In contrast to the BK model, the OFC model is not implemented as a concrete spring-block system nor as a physico-mathematical description of such system. Instead, OFC implement the model as a cellular automaton. Each site *(i,j)* in the cellular automaton is a variable that stands for the tensional forces acting upon block (*i,j)*. The dynamics of the spring-block system is represented by the following algorithm:

Step 1: All sites are initialized with a random value between 0 and F_{th} (the static friction threshold).

Step 2: If any $F_{i,j} \geq F_{th}$ then update the values of i,j and its neighbors in accordance with this rule: $F_{n,n} \rightarrow F_{n,n} + \alpha F_{i,j}$; $F_{i,j} \rightarrow 0$; where $F_{n,n}$ stands for the tensional forces in the neighbors of *i,j*. This step is often referred to as "relaxation." A model-earthquake is evolving.

Step 3: Repeat the second – relaxation – step until all sites have $F_{i,j} < F_{th}$. The number of times this step is repeated amounts to the number of sites being updated. The number of sites being updated – i.e., the size of a chain reaction – is analogous to the amount of energy released in an earthquake.

Step 4: Search the site with highest value, call it F_{max}, and add F_{th}-F_{max} to all sites. Such global perturbation is often referred to as the "driving" of the model. In this case, the driving represents the increase in strain due to tectonic forces at seismic faults. Then return to the second step.

The algorithm swings between driving and relaxation phases. That is, the algorithm stays in the loop between step 2 and 3 before moving to step 4. This separation of driving and relaxation embodies the "separation of timescales" condition. Christensen and Olami (1992a) argue that such separation of timescales in the OFC model mimics the rate of tectonic loading and propagation of an earthquake in actual seismic faults (1832).

As part of its design, the OFC model embodies two features which make it different from most SOC models at the time of its publication. First, it is a continuous cellular automaton. This means that the values attached to the sites – also referred to as state variables – are continuous: they can take any real (positive) number. This feature is distinct from most SOC models, which typically take discrete state variables. Second, the OFC model

is a nonconservative cellular automaton. This means that not all of the tension accumulated in a slipping block is redistributed among its neighbors; part of the elastic potential energy is dissipated. This is an unusual feature among SOC models, as Christensen and Olami (1992a) submit: "A common feature to most of the [SOC] models was that the local dynamical rules obeyed a conservation law" (ibid: 1829). By looking at the definition of α, it is clear that the only case in which there is conservation obtains for $K_L = 0$. However, in order to run the simulation, K_L must be a positive number. Otherwise, the moving plate does not exert any influence upon the blocks. Therefore, the OFC model is nonconservative by design (and so is the BK model).

Note that dissipation occurs in the bulk – due to $K_L > 0$ – but there is also dissipation at the boundary of the fault system. In the OFC case, there is dissipation given that they assume a rigid border condition. This means that force of blocks at the boundary is, by definition, equal to zero. Christensen and Olami (1992a) discuss other boundary conditions, namely free and open boundaries. In the case of free boundaries, blocks at the boundary are connected only to blocks within the fault system. This implies that their coordination number is lower than blocks in the bulk of the fault system. In the case of open boundaries, blocks at the boundary are connected to imaginary boundary blocks, entailing that their coordination number is the same as that of the bulk blocks.

2.3.5 Main Results of the OFC Model

OFC report results of simulations in an isometric setting, i.e., K_1 is equal to K_2.[22] They also assume rigid boundary conditions, i.e., $F_{i,L}$ and $F_{L,j}$ are equal to 0. OFC run various simulations with different values for the elastic parameter (α) and size of the lattice (L). The model exhibits robust SOC behavior, manifested in the power-law of the probability density versus size of model-earthquakes (1247).[23] This behavior is robust in the sense that power-law probability densities, including their exponents, stay the same after modifications in the size of the lattice (L) or the addition of noise (see Figure 2-14). Power-law probability densities

[22] Christensen and Olami (1992a) simulate anisotropic settings.

[23] Size of model-earthquakes is the number of sites being updated in one loop (steps 2 and 3 in OFC's algorithm). The number of sites being updated in an avalanche is assumed to be proportional to energy release in earthquakes. Accordingly, OFC decide to label the abscissa of their plots "Earthquake energy release E".

also obtain over a wide range of values of α (see Figure 2-15). However, the exponents of the power-laws change with different values of α, i.e., they are non-universal. Furthermore, a transition from power-law to exponential regime is observed for values below $\alpha \approx 0.05$ (see Figure 2-16).[24] For the isotropic case in three dimensions (i.e., $K_L = K_1 = K_2$; $\alpha = 0.20$), predicted B-values approximate observed B-values in naturally occurring earthquakes (arrows in abscissa, Figure 2.16).

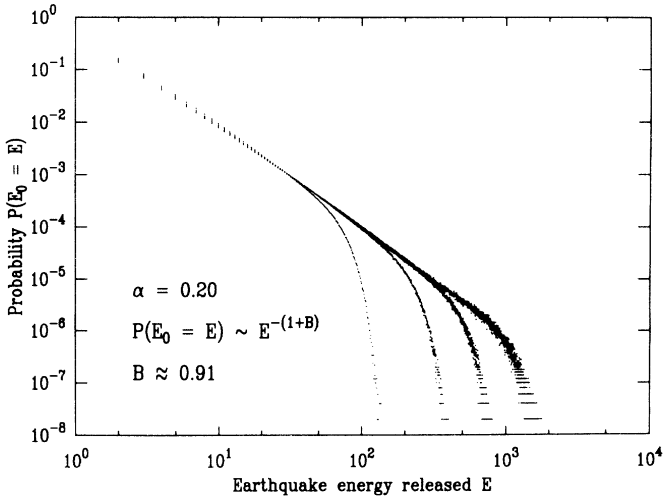


Figure 2-14: Probability of earthquakes according to their energy release. For $\alpha = 0.20$, four different values of lattice's size (L) are tested: L = 15, 25, 35 and 50. A power-law probability density, with exponent B≈0.91, obtains for all values of L. The cut-off in energy distribution scales with $L^{2.2}$ (OFC, 1992).

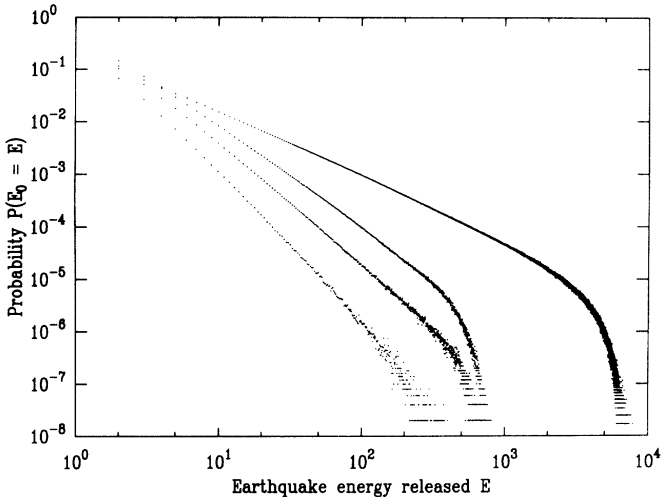


Figure 2-15: Probability of earthquakes according to their energy release. For L = 35, four different values of elastic parameter (α) are tested: $\alpha = 0.25, 0.20, 0.15, 0.10$. The exponents of the curves (B) are inversely proportional to α.

[24] For $\alpha \approx 0.05$, 20% of the tension on the slipping block is redistributed; 80% is dissipated. As a general rule, the amount of conserved energy equals 4α.

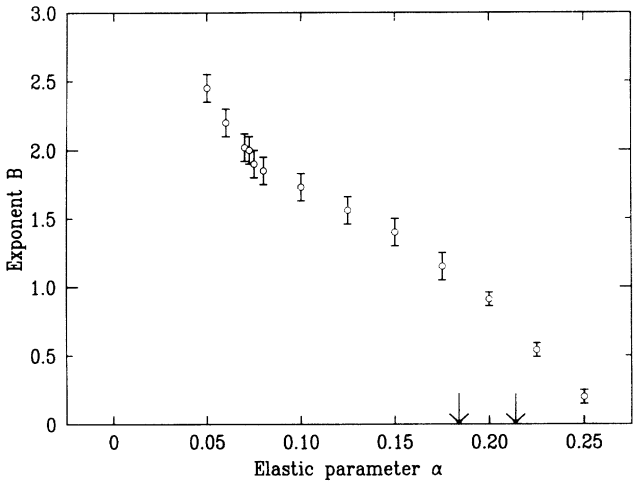


Figure 2-16: Exponent of the power-laws (B) as a function of the elastic parameter (α). Arrows in abscissa indicate B-values for naturally occurring earthquakes (OFC, 1992).

A general conclusion of OFC's research – one that was firstly shown by the BK model – is that there is no distinct cause for small and big earthquakes. Linear intuitions often make us infer that big events have big causes, while small events have small causes. OFC show that the same mechanism produces avalanches of all sizes. This result aligns with central claims in the SOC program.

With this, I close the review of my case studies. In the following, I proceed to develop my philosophical argument. I use my case studies to keep the argument moving forward and exemplify my ideas. As the argument progresses, I present new aspects of the case studies that help me support my views. Still, this chapter contains most of the content necessary to follow my argument. Because of this, I will be referring to content and figures from this chapter regularly.

***

3 Scientific Commitments

Scientists conduct their practices influenced by a series of communal – or even social – guidelines which influence their research. In addition, scientists – as autonomous and creative thinkers – can also introduce original contributions to their practices and follow personal deliberations. I refer to these communal and personal guidelines that shape the practices of scientists collectively as "scientific commitments." BK and OFC have their commitments too, or so I shall argue. In this chapter, I discuss the various commitments that guide BK and OFC's research about earthquakes.

In order to do this, I must set a *minimal framework* for using the concept of 'commitments'. This minimal framework focuses on three main issues: i) an explication of the concept of 'commitments'; ii) a schematic model of the organization of commitments in the context of research programs; iii) strategies for the justification of commitments. The scope of this minimal framework is bound by three considerations. First, the intended context of application is that of my argument, which is based on the examination of two case studies. Second, for the rest of my argument, I will mostly be focusing on one group of scientific commitments, namely explanatory commitments. Third, the framework is mainly stipulative: it states how the notion of "commitments" is meant to be used in this dissertation. Having established the scope of this minimal framework, I must say that I do expect my argument to be sound and useful in other contexts with similar characteristics to those of my case studies. I also conjecture that some parts of my argument are extrapolatable to non-explanatory commitments. And I think my stipulations are well-informed and not widely off the mark. However, this is a matter for further research.

The structure of this chapter is as follows. In the first section, I explicate the concept of 'scientific commitment'. I complement this explication with a brief commentary on the diversity of scientific commitments and exemplifications from my case studies (and other ancillary cases). In the second section, I characterize the organization and dynamics of scientific commitments in research programs. One feature of this organization is particularly salient, namely the differential contribution of commitments to a research program. In the

third section, I discuss the justification of commitments. The main contribution of this section is the presentation of an account of justification of commitments within a research program, to which I refer as coherentism-cum-entrenchment. In each one of these sections, I use my case studies to show how the theoretical discussion bears on them.

3.1 Explication of Scientific Commitments

To begin with, I explicate the concept of 'commitments' in the context of scientific research. This is a crucial task because my overall argument relies upon the notion of "scientific commitments", particularly in the form of "explanatory commitments". To be clear, this explicatory task does not amount to a definition. That is, I do not attempt to say what the necessary and sufficient conditions of something to be called 'commitment' are. Rather, I only attempt to propose a construal of the concept of 'commitment' that is fruitful, simple, and precise enough to be used in my argument (cf. Carnap 1950: 7).

An expedient starting point in any explicatory task is the dictionary. My strategy is to look for the term 'commitment' and discuss its various meanings. More explicitly, I assess whether these meanings cover or do not cover relevant aspects of the concept as it is intended to be used in my argument. Consequently, I allow these relevant aspects to shape my explication of the term. "Lexico" defines 'commitment' as: a) The state or quality of being dedicated to a cause, activity, etc.; b) An engagement or obligation that restricts freedom of action; and c) A pledge or undertaking. These definitions capture relevant aspects of the notion of "commitment", as I intend to use it in the context of my argument. However, none of these definitions does the whole work, and some of them involve ideas from which I would prefer to depart.

First, (a) rightly captures the dedication to a cause (or activity) as part of the notion of "commitment". Scientists dedicated to a same cause share – at least – one commitment, namely the cause. Nonetheless, the way in which I intend to use the term 'commitment' involves more than just being dedicated to a cause: it also involves the way in which scientists are dedicated to that cause. In other words, I intend to use the term 'commitment' to identify ends, but also means. In this sense, scientists dedicated to the same cause might have

different commitments concerning the means to achieve the cause. And scientists with different – even incompatible – causes might share other commitments as means, such as methods or standards.

Second, there are two aspects in (b) that I consider misleading in an explication of 'commitments', namely "obligation" and "restrictions on freedom of action". The term 'obligation' suggests that scientists are bound to follow certain commitments. This is misleading: scientists are active agents in choosing, configuring, modifying, and renouncing to commitments. To be sure, individual scientists are typically accountable to communal norms or standards, but these are negotiable. With regards to "restrictions on freedom of action," I do not fully disagree with the characterization. After all, by following a given set of commitments, the resulting actions span a narrower spectrum of possibilities. This can be interpreted as commitments "restricting the freedom of action." However, I suggest that an explication of 'commitments' should emphasize their role as enablers, as opposed to restrictors. Commitments guide decisions and thus enable actions to be made. Guidance should not be regarded as restriction, but rather as direction.

Third, the explication of 'commitment' as "a pledge or undertaking" in (c) makes us wonder: To whom are scientists pledging their commitments? Who is the recipient – and arbiter – of such pledges? Are the recipients the scientists themselves, their peers, the community at large? Furthermore, in which setting are such pledges explicitly proclaimed? I consider this is the wrong way to go about commitments. By posing commitments as pledges, one is guided to think that commitments are supposed to be honored and fulfilled, even in light of obstacles, as some kind of burden, to respect other agents' integrity or expectations. This is not how I intend to use the notion of "scientific commitments". In my view, commitments can be given up, often readily given up. Thus, I suggest that the right kind of epistemic attitude that an agent has towards commitments is one of acceptance. Acceptance can be interim and contingent, thus departing from notions such as "pacts", "obligations", or "liabilities".

In light of my previous commentary on the definitions of 'commitments', I suggest the following explication: scientific commitments are general guidelines, accepted by working

scientists, which shape their research. Three aspects of this explication are particularly salient, namely i) what kind of thing commitments are; ii) what their relation to scientists is; and iii) what their role in scientific research is. First, commitments are explicated as guidelines. The term 'guidelines' is admittedly vague. However, this is precisely what I want my explication to do: to remain vague about how those guidelines can be embodied. This is because of the varied nature of scientific commitments. Consider for a moment the dictionary's entry on 'guidelines'. "Lexico" defines 'guideline' as "A general rule, principle, or piece of advice". Rules, principles and advices are different kinds of entities, which operate in different ways. Nonetheless, they converge in their ability to guide decisions. This is the crucial aspect about commitments: guidance. Commitments are entities that guide decisions and actions. Thus, I suggest that the vagueness of the term 'guidelines' is a strength of my explication. It sacrifices precision at the expense of scope and fruitfulness.

As a clarification, I highlight an idiosyncratic feature of my explication of 'commitments'. Typically, 'commitment' is taken to be an attitude or disposition of an agent towards a certain guideline, as I have explicated above. That is, scientists are committed *to* a guideline. However, in my explication, I mostly use the term 'scientific commitment' to refer to the object of such attitude or disposition. In this sense, scientific commitments are guidelines. This ambivalent approach to 'commitments' embodies a means-end dichotomy. On the one hand, commitment is a task in which scientists engage. On the other hand, commitments are the consequence of engaging in such task.[25] While I admit the processual character of commitments, my focus is on its objectual character, i.e., commitment as guideline. This focus responds to practical considerations: the target of my study are the guidelines themselves.

The second relevant aspect of my explication of 'commitment' is the relation that scientists establish towards them, namely a relation of "acceptance." Acceptance of a commitment is a disposition to act in accordance with it. For this reason, acceptance involves ability: one cannot accept that what one does not know how to act upon (cf. Elgin, 2017: 19-

[25] Other notions relevant to my argument which also embody a similar means-end dichotomy are "judgment," "interpretation," and "understanding." I discuss these notions in due course.

20). Acceptance often has internal manifestations in the form of conscious willingness to act in accordance to a commitment. However, tacit or implicit acceptance can also occur. In the latter case, acceptance can be assessed externally. I do not attempt to decide whether an internalist or externalist approach is a better strategy to characterize acceptance as an epistemic attitude. I only acknowledge the existence of these two manifestations.

Acceptance is a matter of degree: one could be more or less committed to a certain guideline. The degree of acceptance depends on the abilities of the scientist and on the scientist's assessment of the relations among her other commitments. Sometimes, this assessment may be based on consideration of certainty or accuracy. However, it is worth emphasizing that acceptance of a commitment does not (necessarily) imply a commitment being true, nor believing the commitment to be true, nor even the commitment being truth-apt.

I submit that acceptance of commitments occurs at the level of individual scientists, i.e., at the personal level. However, as a figure of speech, I also use the term 'acceptance' in the context of groups of researchers or communities. A group of scientists can be said to accept a commitment insofar as the number of individual scientists accepting the commitment surpasses an agreed threshold-criterion or statistical consideration. The claim that commitments are accepted at the personal level does not mean that they are entirely subjective. After all, commitments are intended to direct actions in the domain of scientific research, which has intersubjective norms. In this sense, scientific commitments are often contextual: one same scientist could accept different – even incompatible – commitments in different scientific contexts.

The third relevant aspect of scientific commitments is that they shape scientific research. I use the term 'shape' to depart from the idea of commitments "determining" the outcomes of research (or other similar notions). In other words, commitments underdetermine the developments in a research program. Commitments can be regarded as dispositions. But – as dispositions – commitments operate within a constellation of other commitments which allows for multiple possible outcomes. This is the case because

commitments can be weighed up, combined, and relaxed in multiple ways (cf. McMullin 1982: 17).

Here, I find it useful to introduce the notion of "judgment". The concept of 'judgment' can be used in two distinct ways, namely as a process or as an outcome. As a process, judgment is the ability to make decisions. As an outcome, judgment is the content of such decisions. The effect of commitments on judgments can be seen in both the process of making decisions and the decisions themselves. But the effect of any commitment in a decision-making process depends on its relation to other commitments in specific contexts. In other words, judgments do not follow univocally from the acceptance of a certain set of commitments: how those commitments are organized is relevant to the judgments. In fact, there are rational discussions to be had on the various judgments that may be made based on a same set of commitments.

After delivering an explication of 'commitments' in the context of scientific research, I proceed to deliver a brief commentary on the variety of commitments. To be sure, I do not have any expectation of exhausting the topic. Rather, my intention is to provide the reader with a "taste" of the range of categories and heterogeneity among scientific commitments. To begin with, scientific commitments can be embodied in various kinds of entities. For example: values, rules, exemplars, attitudes, judgments, principles, practices, advices, norms, concepts, methods, customs, heuristics, standards, beliefs, and techniques, just to name a few. (These entities are not necessarily mutually exclusive.) In terms of their content, commitments are widely heterogenous across scientific disciplines. To be fair, there are some ubiquitous and rather stable commitments, such as empirical adequacy, logical consistency, and several mathematical principles. But there are plenty of scientific commitments that are more idiosyncratic to specific research programs, such as specific working hypotheses or the methods employed in particular research groups.

Scientific commitments can be of various sorts. Although I focus on commitments of an epistemic nature in this dissertation (namely explanatory commitments), it is worth mentioning that scientific commitments can cover metaphysical, ontological, ethical, institutional, technical, political, aesthetical, and cultural considerations. Certainly, a

discussion on the various kinds of commitments should be more nuanced. For matters of space, I do not engage in this discussion. However, I point at two issues that should be addressed in such discussion. First, it is not straightforward to establish categories of commitments. As an illustration, consider the distinction between epistemic and non-epistemic commitments (see e.g., Reiss & Sprenger, 2014). Various philosophers have made the case that these categories are intricately intertwined (see, e.g., Douglas, 2016). Philosophers arguing for this position go from advocates of standpoint theory to sceptics of the distinction between context of discovery and context of justification. Second, even if kinds of commitments are well-established, it is not straightforward to classify specific commitments into one or another category. For example, simplicity has been discussed as an epistemic commitment, but also as a non-epistemic (e.g., aesthetical) commitment.[26]

Commitments can vary in their degree of explicitness. I highlight two contexts in which this is the case. First, explicit acceptance of a commitment might come with an implicit acceptance of subsidiary or more basic commitments. For example, by explicitly accepting a method, one is implicitly accepting commitments related to the method, e.g., ontological commitments that the method requires. Second, acceptance of commitments might become customary and thus unreflective and implicit. For example, this is the case with standard procedures or protocols which guide scientists' actions in a way that becomes customary. Scientists do not need to be fully aware of their commitments to act in accordance with them. But then a problem arises: Tacit guidance might be confounded with coincidences. That is, one might decide something or act in accordance with a commitment but only by chance. I suspect that there are criteria to distinguish tacit guidance from coincidences, but I do not engage fully in this line of research. Having said this, I do discuss tacit knowledge in Chapter 7.

[26] Given the widely diverse nature of commitments, it is relevant to comment on what does not count as a commitment. I suggest two restrictions. First, scientists' personal experiences such as pervasive emotions and sensations are typically not commitments. These items lack the affirmation required for commitment and they are not – at least, not straightforwardly – connected to research decisions. Second, commitments in one context might not be commitments in another context. Commitments involve acceptance, and acceptance of specific guidelines is contextual.

As an illustration of scientific commitments, consider the BK case. A good example can be identified in the very title of BK's paper: "Model and Theoretical Seismicity." The title is a reference to model and theoretical methodological approaches in seismology, which BK intend to replicate in the context of the problems of seismicity. The distinction between these two approaches is spelled out in terms of distinct "underlying principles" (BK: 341). I submit that these distinct "underlying principles" that shape the theoretical and model approaches are instances of scientific commitments: they are accepted guidelines that shape BK's research decisions.

These commitments can be characterized as methodological commitments: they shape methodological decisions to deal with a research problem. BK's paper reports the employment of both methods. The theoretical approach to seismicity is displayed in BK's physico-mathematical description of the generalized spring-block system, in the form of the equations of motion of the blocks. The model approach to seismicity is displayed in BK's computer model that solves the physico-mathematical description using a numerical model. In this case, a computer program solves the equations of motion – a set of first order differential equations – by using a Runge-Kutta's procedure.

After explicating the notion of scientific commitments, delivering further characterizations and illustrations, and showing how this explication bears on my case studies, now I move on to a more fine-grained characterization of the organization of scientific commitments in research programs.

3.2 Organization of Scientific Commitments

Before engaging in scientific research, scientists are initially inclined to accept some commitments. These initially tenable commitments are the best take that scientists have on the matter under research in an early stage. As Elgin (2017) argues – paraphrasing Quine – inquiry always begins in *medias res*: scientists stand "somewhere" with a particular perspective (64). Such perspectives involve commitments of various sorts. Accordingly, Elgin rejects the possibility of a "view from nowhere," a "God's eye point of view," or an "innocent eye" in scientific research (cf. Giere 2006; Gombrich, 1960).

These constellations of initially tenable commitments might change as research unfolds. But the changes are not arbitrary. They aim for the improvement of the constellation of commitments. Elgin (2017) argues that the improvement of a constellation of commitments amounts to the strengthening and widening of the relations of mutual support among the commitments. By rejecting, correcting, revising, and augmenting commitments, scientists bring the collection of their commitments into accord.

The state of (meta)stable mutual accord among commitments is what Elgin refers to as the state of "reflective equilibrium" (cf. Rawls, 1999). In reflective equilibrium, accepted commitments are reasonable in light of one another (this is the "equilibrium" component). In addition, the constellation of accepted commitments – as a whole – is as reasonable as any available alternative in light of the relevant antecedent commitments. In other words, it is a constellation that the scientific community in question can, on reflection, endorse (this is the "reflective" component).

Reflective equilibrium is not a final state: it is provisional and dialectical. New findings often disturb a state of reflective equilibrium. In this sense, reflective equilibrium does not afford ultimate truth, nor even approximate truth. So much for truth. As Elgin argues, truth should not be construed as the aim of science. Rather, aiming for reflective equilibrium is a better characterization of the scientific enterprise. Reflective equilibrium among commitments is the best scientists can do and should be enough for acceptability of a constellation of commitments. In other words, reflective equilibrium is not an optimum. Rather, it is a suboptimal solution to the problem of fit among commitments.

Elgin (2017) – who has influenced much of my views – refers to these constellations of commitments in mutual support as "accounts".[27] I consider this term a fortunate one, for reasons that will become clearer in the next chapter. The gist of the matter is that the same term – 'account' – is often used to refer to different construals of the notion of explanation,

[27] Elgin (1996) uses the term 'systems of thought' to refer to virtually the same notion as 'account'. In her 2017's book, she prefers to use the term 'account' because – as she argues – this enables her to "avoid getting involved in disputes about the relation between theories, evidence, and models" (311).

known as “accounts of explanations”. I suggest that accounts of explanation are, precisely, (meta)stable bundles of explanatory commitments that scientists may or may not endorse.[28]

Accounts are a basic level of organization of commitments. A higher level of organization of commitments is the level of a “research program.” For the purposes of this dissertation, a research program is a stable collection of commitments accepted by a community of collaborative scientists, which guides their research. The collaboration among scientists is shaped by two main considerations: i) scientists in a same research program collaborate by tackling a common scientific problem and exploring interrelated research questions; ii) scientists in a same research program collaborate by sharing commitments of various sorts, in varying degrees.

Even though a research program is supposed to be stable, it is not static. Scientists actively shape research programs as epistemic agents. I follow Elgin (2017), who claims that scientists are legislators in a realm of epistemic ends (120). As legislators, scientists can only accept commitments that they could advocate and reflectively endorse. I share Elgin’s characterization of this “legislating” process as pragmatic. That is, the commitments that a community holds might evolve relative to the contingent interests of the community. This evolution proceeds in accordance with basic principles of responsibility and self-discipline. Certainly, cases of indiscipline and irresponsibility are not unheard of in scientific research. However, even in these cases, collective regulation can validate or invalidate some of the biased endorsements of commitments. This collective regulation leaves sufficient range for divergence of specific commitments under the same global scientific commitments, i.e., scientists tend to adhere to open-mindedness.

A crucial feature of the various scientific commitments that belong to a research program is that they do not share the same status. This difference in status manifests qualitatively in two distinct ways. First, there are some commitments that scientists are less

[28] Ankeny and Leonelli’s notion of “repertoires” (2016) is similar to the notion of accounts construed as a metastable bundle of commitments. However, repertoires seem to go have a broader scope than accounts, given that they include those material and institutional conditions under which scientists operate and are not willingly or freely accepted. I suggest that these conditions should be understood as “border conditions” for the acceptance of commitments.

willing to modify or renounce to, while there are other commitments that scientists are more prone to sacrifice. Second, there are some commitments that play a more fundamental and ubiquitous contribution in decisions made by scientists during research. In particular, there are commitments that are systematically used as standards to evaluate the acceptance of other commitments. Although these two manifestations are conceptually distinct, they typically cooccur in practice. That is, those commitments that scientists are less willing to give up are those that commonly contribute more fundamentally and ubiquitously to the program. In contrast, those commitments that scientists are more prone to give up are those that do not contribute fundamentally and ubiquitously to the program and tend to be derivative on standard ones.[29]

To capture these qualitative differences, I introduce a conceptual distinction between "core" and "peripheral" commitments in a research program. Core commitments are those that scientists are less willing to give up and play a more fundamental and ubiquitous contribution to decision-making in the program. Peripheral commitments are those that scientists are more prone to give up and are either derived, evaluated, or judged by means of core commitments. This distinction between core and peripheral commitments entails a binary structure within research programs. That is, research programs have a core and a periphery. The core of a research program is the set of its core commitments. The periphery of a research program is the set of its peripherical commitments. It is important to note that the distinction between core and peripheral commitments does not overrule the requirement for reflective equilibrium. Rather, the distinction only submits that, in achieving reflective equilibrium, commitments' contribution to such equilibrium is unequal.

The distinction between core and periphery is not based on intrinsic features of the commitments, but rather on scientists' attitude towards them and their interrelations. This, however, does not mean that core commitments do not have important features. A significant feature of core commitments – as I intend to characterize them – is that they possess basic semantic and epistemic roles. The semantic role of core commitments amounts

[29] van Fraassen (1980) submits that acceptance of theories can be full, tentative, or to a degree (12). I suggest that the same applies to the acceptance of scientific commitments.

to their disposition to afford meaning. Such meaning can be attributed to words, sentences, images, objects, actions, to name a few entities. In particular, I am interested in how core commitments are used to provide meaning to scientific models (see "interpretation" in Chapter 5). The epistemic role of core commitments amounts to their disposition to afford reasons that promote certain beliefs. This is particularly relevant in the crafting of explanations: core commitments afford reasons to believe explanatory content derived from models (see "explanatory commitments" in Chapter 4). Note that the semantic role is more fundamental than the epistemic one: meaning is necessary to rationally believe a claim. This implies that the interpretation of models is prior to the crafting of model explanations (see Chapter 5). To be sure, semantic and epistemic roles are not exclusive to core commitments. Peripheral commitments might have them as well. However, the scope of these roles in peripheral commitments is more restricted.

The distinction between core and periphery in research programs is – I admit – questionable on various fronts. I mention four potential critiques. First, my views suggest that commitments can be univocally assessed in terms of their contribution to research programs and thus be assigned to the core or the periphery. Nonetheless, the contribution of commitments can be assessed along various dimensions or criteria. How to integrate these various assessments is far from clear. In fact, how the integration of various assessments proceeds depends itself on other, higher-order, commitments. This brings us to the second problem: the holistic character of commitments. The value of commitments cannot be assessed in isolation or atomistically. Commitments relate to each other in complex ways that make the assessment of the impact of single commitments impractical. Third, even if a straightforward assessment of the contribution of commitments was attainable, it is not clear how to draw the line between core and periphery. In other words, although the conceptual distinction between core and periphery is clear, it is often difficult to establish such distinction in practice. Arguably, the best way to visualize the distinction between core and periphery is not as a sharp boundary between two sets of commitments. Rather, the core and the periphery should be regarded as two regions with a transitional zone. In other words, I suggest

that being in the core or the periphery is a matter of degree.[30] Fourth, commitments' contribution to a research program typically change as research unfolds within a program. This makes the core and periphery dynamic entities.

Although I recognize the legitimacy of the aforementioned critiques, they miss a crucial point. The core-periphery distinction might be difficult to establish in practice. But the purpose of introducing the distinction is not a classificatory one. Rather, the distinction is intended as an operational distinction that allows us to talk about a widely reported feature of scientific practice. Such feature is that commitments that guide scientific practice contribute differentially to the practices conducted within a program. That is, some commitments are prior to others in decision-making, or have a more ubiquitous influence across decisions, or scientists are less willing to renounce them.[31]

To visualize the relations between accounts and research programs, I provide a schematic representation (see Figure 3-1). The periphery of the research program is depicted in yellow and it transitions continuously towards the core in red. The various commitments that belong to the program are organized in (meta)stable units, namely accounts, depicted as blue ellipsoids. (Commitments within each account are not depicted.) The various accounts that are part of the research program relate to each other, securing a stable (although not static) program. Some accounts are more peripheral, other are closer to the core. Others are partly in the periphery, partly in the core. This is intended to depict the differential contribution of commitments in endeavors within a program. Note that accounts vary in size which is a proxy for the number of commitments that they involve. Accounts may also exhibit overlaps with other accounts. The boundaries of accounts and the program as a whole are not sharp (hence the soft borders in Figure 3-1).

[30] Dascal (2003), in the context of discussing "pragmatics", makes a useful distinction between commitment and involvement. For him, 'commitment' is a "yes/no" concept, while 'involvement' is a degree concept. For matters of conceptual economy, I prefer to talk about gradual commitments.

[31] Instances of the differential contribution of commitments in research programs are found in the context of alternative construals of scientific research. For example, there is differential contribution of: i) rules in Carnap's linguistic frameworks (1950); ii) exemplars in Kuhn's paradigms (1996); iii) theses in Lakatos's research programs (1978); iv) the mathematical-mechanical-empirical components of theories in Friedman (2001); v) content in Pincock's mathematical representations (2012); and vi) rules in Mantzavinos's explanatory games (2016). An exhaustive discussion of these findings was left out in the final version of this dissertation.

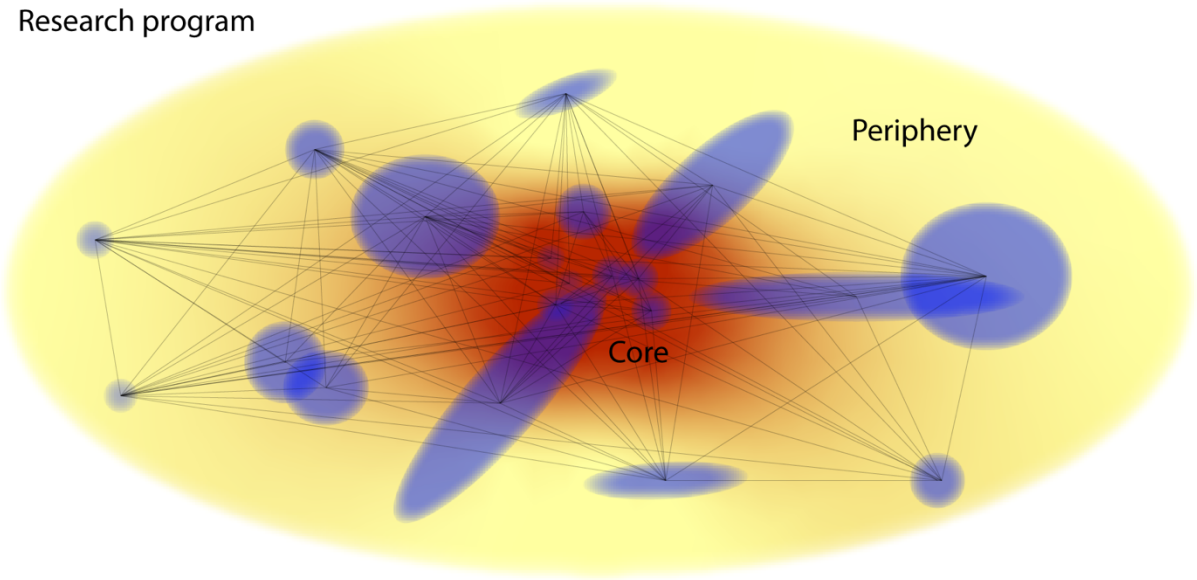


Figure 3-1: Representation of a research program and its constitutive accounts. Accounts are depicted in blue. The core of the program is depicted in red. It transitions continuously towards the periphery, depicted in yellow. Note that accounts are dissimilar in size (number of component commitments) and position within the program (differential contribution of accounts). Also, some accounts may possess commitments which play more or less central roles (see elongated accounts that point towards the core).

To illustrate the operation of the core-periphery distinction in a research program, I single out three commitments adopted by BK in the form of principles. (In the next chapter, I assess the organization of these principles within accounts of explanation.) The first principle is the suitability of a discrete system to represent and simulate the behavior of a continuum medium (P1). The second principle is the sufficiency of friction as a causal factor to account for the main features of the statistics of naturally occurring earthquakes (P2). The third principle is the suitability of the Runge-Kutta procedure to provide numerical results (P3). BK discuss the employment of these principles in their paper at length:

> ***[P1]:*** *"The treatment of problems involving friction leads to descriptions of nonlinear phenomena. To describe the mechanics of continuous media we solve a set of partial differential equations. The solutions to nonlinear partial differential equations are in general very difficult to obtain. One technique that has been applied is to describe the continuum in terms of a set of discrete particles; the corresponding differential equations then reduce to ordinary nonlinear differential equations. We can thus imagine the particles on opposite sides of the fault reduced to a two-dimensional network of masses interconnected by springs which represent the usual elastic elements and which are further*

coupled by frictional elements interconnecting the elements on opposite sides of the fault. Relaxation methods can then, in principle, be applied to solve this set of ordinary differential equations. This is a stupendous problem in general and one which has not yet been attempted. In order to learn something about the process of solving these equations we have recourse to further reducing the set of particles to a one-dimensional array. The numerical calculations that result from this simplification are still overwhelming but not impossible of solution. Thus we consider a one-dimensional analog of the two-dimensional problem of the interrelationship between particles on opposite sides of a fault surface" (342).

[P2]: *"The feature we choose to isolate is the conjecture that a gross form of friction between the two walls of an earthquake fault inhibits the relative displacement of the material on the two sides. Assume that differential driving stresses are present due, for example, to large-scale, slow deformation. When these differential driving stresses exceed the limiting frictional stresses, then the displacement that takes place is sudden, and an earthquake is presumed to have occurred. Following some relative motions between the faces of the fault, the friction then brings the two faces of the fault to rest; the stresses are now less than those corresponding to the "limiting friction", and all is quiet. The relief of stress is assumed to account for the quiet interval between earthquakes; the driving stresses must then be built up once again to exceed the limiting frictional stresses which inhibit the motion" (342).*

[P3]: *"[The set of equations of motion of the spring-block system] is a system of 2N first order differential equations for x_j and y_j and may be integrated numerically, for instance, by a Runge-Kutta method provided due care is taken at the points where the forms of the expressions for $F_j(y_j)$ change" (354). "The entire problem of solving the nonlinear differential equations was left to numerical integration using a Runge-Kutta procedure. The problem was programmed for solution upon an IBM 7094 computer" (359).*

I submit that these principles do not share the same status. P1 and P2 play a central and ubiquitous role in the overall design of the BK model. In other words, these are core commitments. This is expressed early on in BK's paper:

> *"In this paper we wish to isolate one of the qualitative features used in one of the models for earthquake occurrence [namely friction, i.e., P2]. We shall construct a mathematical description of the problem and also construct a highly simplified laboratory model of this problem [a discrete model to represent a continuum, i.e., P1]. We then proceed to see how many of the features observed in nature can be simulated by it" (342).*

In contrast, P3 plays an important role but one that is restricted to a later stage of BK's study. More specifically, P3 is not employed in the design of the BK model but only used to obtain numerical results from the physico-mathematical description of the spring-block model. Furthermore, P3's role in BK's study can be fulfilled by alternative methods without changing the overall contributions of the study. This is actually hinted at by BK, who claim that the system of differential equations "*may* be integrated numerically, *for instance*, by a Runge-Kutta method […]" (354; my emphasis).

The difference between the peripheral status of P3 versus the central status of P1 and P2 is manifested in BK's different epistemic attitudes towards these principles. P1 and P2 are principles that, by being implemented in the BK model, are intended to be tested and demonstrated via the simulations. In this sense, the BK model can be characterized as a proof-of-principle for P1 and P2. For P1, the proof-of-principle is slightly different in the two implementations of the BK model. In the laboratory implementation, the proof-of-principle consists in constructing a discrete mechanism that simulates the statistics of occurrence of earthquakes in (continuous) geological faults. In the mathematical implementation, the proof-of-principle consists in showing that the solution of a set of ordinary nonlinear differential equations approximates that of a set of partial nonlinear differential equations. For P2, the proof-of-principle consists in isolating friction as a causal factor in both the laboratory and mathematical implementations of the BK model. Then, simulations are run, and the results are compared to statistical data of naturally occurring earthquakes. If the simulations' results resemble the behavior of naturally occurring earthquakes, then P2 is demonstrated. For P3, things are quite different. P3 is not part of BK's research agenda. In other words, P3 is not implemented to be tested or demonstrated. Quite the opposite. BK implement P3, a widely used efficient method, to solve a problem, namely deriving results

from a system of nonlinear differential equations. Furthermore, this problem could be solved by using alternative principles.

A reasonable critique to my analysis is that there might be other commitments even more fundamental and ubiquitous than P1 and P2. Think of, for example, basic mathematical theorems, or – for that matter – grammar rules in English language. If one concedes this point, then one might argue that the core of BK's program can be downsized to the extension of these more basic commitments, leaving P1 and P2 in the periphery. I admit that mathematical and natural language commitments play a role in BK's work. After all, how could BK construct their model without being committed to mathematical theorems or even rules of English language? Having said this, I posit that there is nothing idiosyncratic about these commitments in BK's research. These commitments are shared across widely dissimilar research programs. They are – one could say – highly entrenched commitments, cutting across scientific disciplines, operating in the background.

Thus, the way in which I intend to use the notion of core is as an element that gives a research program its identity, making it different from other research programs. In this sense, highly entrenched commitments are not synonymous with core commitments. I concede that the identity of a research program is dynamic and at any moment it might be extremely difficult and controversial to decide which are its main features. However, I also submit that there are platitudes that drive research programs forward and make them distinct from other scientific endeavors.

In the BK case, BK's theoretical and model seismicity is a research program distinct from the theoretical and model seismology of their days (even if they share mathematical theorems and the English language). I suggest that what distinguishes BK's program from traditional seismology – i.e., what makes it idiosyncratic – is precisely their core commitment, namely (at least) P1 and P2. These principles are BK's matter of study and, accordingly, shape the design of the BK model fundamentally. In addition, P1 and P2 were not (at the time) at the core of seismology, at least not as characterized by BK. That is, seismology was not concerned with the topic of production of earthquakes but rather on the propagation of

seismic waves (i.e., no P2). And seismology dealt with seismic wave equations by using partial differential equations, assuming the propagation in a continuum (i.e., no P1).

Additionally, I suggest that it is the conjunction of P1 and P2 what contributes to the idiosyncrasy of BK's program. More explicitly, other research programs might study the role of friction in the production of earthquakes without employing discrete representations of a continuum (i.e., P2 without P1). Or other research programs might study the suitability of a discrete system in representing and simulating a continuum without having anything to do with the study of friction in geological faults (i.e., P1 without P2). These hypothetical programs are not identical to BK's program. Naturally, this does not prevent these hypothetical programs to inform and contribute to BK's program.

Finally, note that P3 does not contribute to BK's idiosyncrasy. The Runge-Kutta method is shared by several research programs which employ it without studying it. Furthermore, BK's program might have come to a completion by resorting to alternative principles. To be clear, the Runge-Kutta method might be a core commitment in other research programs. This is the case within research programs in mathematics which study the method, e.g., programs that search for proofs or optimizations for the Runge-Kutta procedure.

I conclude that P1 and P2 are core commitments because they play a ubiquitous role in the design of the BK model and in framing BK's research. In addition, BK's study is intended to demonstrate the validity of these principles. In this sense, P1 and P2 are idiosyncratic to BK's program. P3 is a peripheral commitment, which plays a local role in BK's research and it can be replaced by alternative commitments without affecting BK's overall program. Thus, having presented the notion of research programs and the differential contribution of commitments within them, now I proceed to discuss the justification of commitments in research programs.

3.3 Justification of Scientific Commitments in Research Programs

In this section, I attempt to deliver a scheme for the justification of scientific commitments. This is a challenging task. Justification is a broad and often complicated topic. Indeed, it has been the subject matter of a myriad of papers and books in epistemology and philosophy of science. And it has been addressed from different and nuanced perspectives, which are difficult to conciliate, and often simply difficult to understand. Therefore, I consider it valuable to declare what I am not trying to do in this section. First, I do not attempt to deliver a comprehensive review of accounts of justification. I do discuss what I deem to be platitudes in the literature but soon channel them to topics relevant to my own argument. Second, I do not attempt to deliver a general account of justification, nor have a discussion in abstract terms as epistemologists tend to do. My motivations for discussing the justification of scientific commitments are less theoretical and more practical: scientists are asked – by their peers and other groups – to provide justification for their commitments. Third, the scope of my discussion on justification is, primarily, that of my case studies. I suspect that my final views on the topic might have applications in other contexts. However, this is a matter for further research.

Given the practical motivations of this section, I frame my discussion on the justification of commitments in terms of two distinct contexts in which scientists typically find themselves. The first context is that of scientists engaged in a research program, following commitments that are internal to the program. I call this the programmatic context of justification. The second context is that of scientists who accept commitments that are external to their research programs, commitments that challenge the core of the program. In other words, these external commitments are not initially coherent – in any obvious sense – with the core of the framing program. I call this the prospective context of justification.

My thesis is that these two contexts call for distinct approaches to justification. In the programmatic context, I suggest that a mixture of epistemic and practical justification is in place in a form of coherentism with entrenchment. In the prospective context, I suggest that a practical justification is in place, which might have an immediate negative effect in the overall coherence of the research program. The practical justification consists in exploring alternative configurations of commitments in a research program, which are promising solutions to long-standing problems and questions. In this section, I address the justification

of commitments in the programmatic context. I leave the justification of commitments in the prospective context for Chapter 8, where I address the issue of exploration.

3.3.1 Epistemic and Practical Justification of Commitments

Justification can be characterized as the task – or process, if iterated – of providing good reasons. Depending on the kind of reason, two distinct kinds of justification can be recognized, namely epistemic and practical. This distinction has been mostly discussed in the context of justification of beliefs (e.g., see Sosa, 1980). Nonetheless, I suggest that this distinction holds for the justification of commitments as well. For purposes of clarity, I present the distinction in the context of justification of beliefs. Later, I proceed to use the distinction in the context of justification of commitments.

The distinction identifies two kinds of justification, based on the kind of reasons being used to justify. On the one hand, there is epistemic (or theoretical) justification: a belief is epistemically justified if the reasons employed in the justification are pieces of evidence and/or bits of knowledge. On the other hand, there is practical justification: a belief is practically justified if having the belief advances the achievement of goals. The following example, provided by Sosa (1980), best illustrates the distinction: “Someone seriously ill may have two sorts of justification for believing he will recover: the practical justification that derives from the contribution such belief will make to his recovery and the theoretical justification provided by the lab results, the doctor’s diagnosis and prognosis, and so on” (ibid: 3).

The justification of commitments is the task of providing good reasons in favor of accepting those commitments. I submit that these reasons can also be epistemic and/or practical, i.e., bits of knowledge and/or motivations. This means that scientists might accept a commitment based on their knowledge. Or, they might accept a commitment based on its role in advancing her goals. Note that these reasons – i.e., bits of knowledge and motivations – are themselves commitments.

As an illustration, consider BK's commitment to P1 and P2 introduced above. Recall that P1 is the suitability of a discrete system to represent and simulate the behavior of a continuum medium. And P2 is the sufficiency of friction as a causal factor to account for the main features of the statistics of naturally occurring earthquakes. I suggest that P1 and P2 are justified differently. P1 is justified practically. P1 is not justified based on knowledge of geological faults. BK know that geological faults are not discrete systems (at least, not in the sense they are represented in the BK model). BK are justified in accepting P1 because it is a convenient technique that affords a solution to a problem. In fact, P1 is a "highly simplified" way to represent a geological fault (BK: 342). But because of this, P1 affords intelligibility and manipulability to the BK model, which are clear desiderata in BK's research.

P2 is justified in a more hybrid fashion. On the one hand, P2 is justified epistemically. After all, BK know that friction is an important causal factor in the occurrence of earthquakes, thanks to qualitative theories available at the time.[32] On the other hand, P2 is justified practically. It is convenient, for matters of tractability and controlled manipulation, to isolate friction as a causal factor. This is evidenced by a whole section on BK's paper focused on the design of a realistic law of friction for the physico-mathematical description of the spring-block model.

3.3.2 Agrippa's Trilemma in the Justification of Commitments

The evaluation of good reasons is not restricted to the assessment of the content of singular reasons – epistemic or practical – in support of a given commitment. In addition, the evaluation of reasons also involves a broader inspection into the overall structure of the relations between the various reasons provided in a justification process. More explicitly, by means of systematically questioning the legitimacy or adequateness of a reason, more reasons come into play, generating a process of justification. The ways in which these reasons relate to each other are susceptible of evaluation.

[32] As an illustration, consider the following quote: "A number of qualitative theories have been constructed which describe the earthquake focal mechanism and the interrelationship between successive earthquakes in time and space. These theories have perforce been made elaborate because of the large number of features they have been called upon to explain. However, these have remained qualitative theories. In this paper we wish to isolate one of the qualitative features used in one of the models [i.e., qualitative theories above] for earthquake occurrence [namely friction]" (BK: 342)

Given that commitments are justified by reasons which are, themselves, other commitments, the demand for further justification soon collapses into an "infinite regress." That is, commitments that justify commitments require other commitments that justify them in turn *ad infinitum*. To avoid infinite regress, two traditional alternatives are available. First, there is "foundationalism," which posits that some commitments are justified independently of other commitments. With these foundational commitments, a justificatory process comes to an end, i.e., infinite regress is halted. Second, there is "coherentism," which posits that commitments do not need to be justified via an infinite sequence of commitments. Rather, commitments can be justified by fitting into a finite set of commitments standing in coherent relations with each other. Unfortunately, foundationalism and coherentism have problems of their own. Thus, one seems to be left to choose the least bad among three flawed alternatives.

The aforementioned predicament is analogous to a well-known problem in epistemology, commonly referred to as "Agrippa's (or Münchhausen) Trilemma" (Figure 3-2). The trilemma was originally articulated – and has been mostly discussed – as a predicament in the justification of beliefs. Still, I suggest that lessons from discussing this trilemma in the context of justification of beliefs can enlighten my discussion on the justification of commitments. The main lesson I take is the following: one way or another, either by falling into infinite regress, coherentism or foundationalism, processes of justification fail to convince the sceptic. Epistemologists still discuss this issue with the hope of finding a solution. They often provide original and nuanced versions of infinite regress, coherentism and foundationalism that seemingly overcome some of its pitfalls. Still, there is not consensus on its solution.

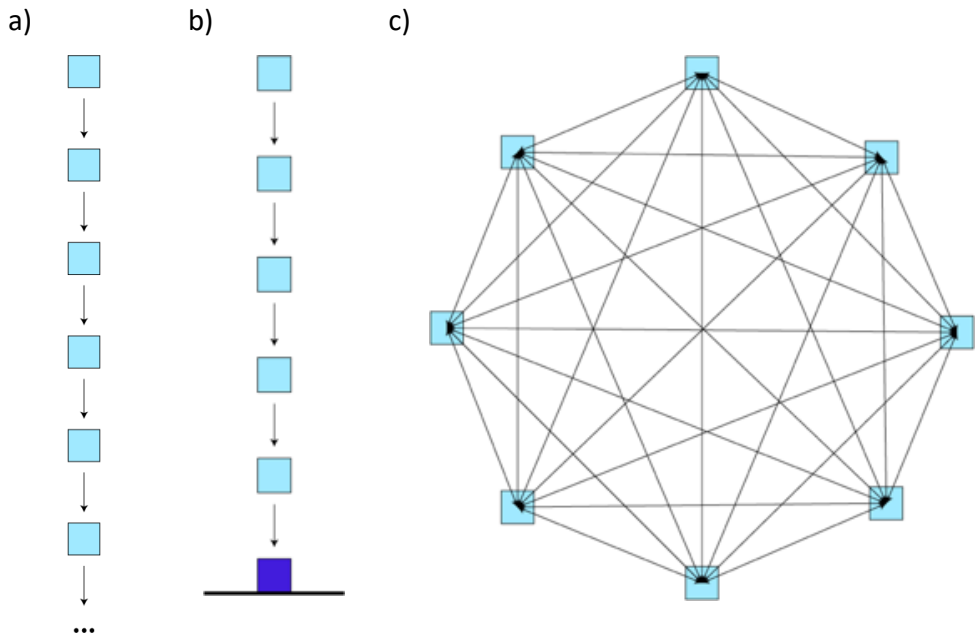


Figure 3-2: Schematic representation of the three alternatives of Agrippa's trilemma: a) infinite regress; b) foundationalism; and c) coherentism.

Given these difficulties, my aims are rather humble. I do not intend to provide a solution to the trilemma. Several efforts towards that goal have been delivered (and mostly failed). I am not in a better position to attempt such an enterprise. Instead, I provide an account of justification of commitments which is informed by this literature but – ultimately – is susceptible to the criticism of epistemologists. Still, it is not an aim of this dissertation to deliver a solution that leaves the sceptic epistemologist content. Rather, I attempt to deliver an account of justification of scientific commitments in research programs that simply reflects – even if partially and fallibly – the attitudes of scientists engaged in such programs.

The difference between these two projects should be clear: scientists are not, as opposed to epistemologists, in the business of developing – or acting in accordance with – a general and abstract theory of justification. Scientists' approach to justification is more practical and contextual, based on the concrete problems that they face. And the actual justification of commitments that scientists provide need not fulfil epistemological standards of what justification is *in abstracto*. In this sense, the nature of my account of justification of scientific commitments has less to do with the picture that epistemologists have in mind and more to do with that of philosophers of science in practice.

3.3.3 Justification of Commitments in Research Programs: Epistemic and Practical Justification Based on Coherentism-cum-Entrenchment

Hoping to have clarified the nature and scope of my account, I proceed to present it. There are two basic tenets that are relevant to my account at which I already have hinted above. The first one is that justification of commitments can be epistemic or practical. The second tenet is that commitments are justified by other commitments. In claiming that scientific commitments are justified by other commitments, I can follow either the infinite regress path or the coherentist path. Foundationalist approaches fail because commitments cannot be justified independently of other commitments. In other words, there is no such thing as a foundational commitment.

From the two available alternatives, I discard the infinite regress path as a viable option, because it fails to do justice to actual justificatory practices. Scientists in research programs do not engage in endless chains of justification to support their commitments. There is a sense in which – one could argue – there is an endless process of justification of scientific commitments unfolding in time or, more precisely, historically. Current scientific commitments are justified by earlier commitments, and the latter are justified by even earlier commitments, and so on. In practice, however, these are not the kind of justificatory process in which working scientists engage.

This leaves me with the "coherence" alternative. My research suggests that scientists justify their commitments in relation to a finite set of other commitments that form part of their research program and its constitutive accounts. In other words, a coherentist approach does more justice to scientists' practices of justification of commitments. The coherentist view I have in mind aligns with Elgin's approach to scientific commitments (2017). In her view, scientists aim for a sort of reflective equilibrium across the constellation of their accepted commitments. Such reflective equilibrium amounts to the mutual impact that commitments have in shaping each other and reaching an overall (meta-)stable configuration. Reflective equilibrium brings about coherence across the constellation of commitments.

I am interested in assessing coherence at the levels of accounts and the level of research programs. In contrast, I admit less than coherent configurations at the levels of scientific disciplines or individual scientists. Coherence among commitments in an account or research program is often not complete. That is, there might be commitments within an account or research program which conflict with other parts of the account or program. It is not straightforward to indicate the threshold of coherence that an account or research program has to minimally achieve. Arguably, this can only be responded contextually. Still, there is a tendency towards coherence driven by processes of reflective equilibrium. In this sense, some incoherencies within an account or research program might be tolerable and taken to be contingent and temporary.

Justification via coherentism comes in degrees. That is, scientists are more or less justified in accepting a certain commitment in the context of their adopted accounts and home research program. Instead of a problem, I suggest that the partiality of the justification of commitments aligns well with the partiality of acceptance of commitments. Given that the acceptance of commitments is a matter of degree, their justification does not need to hold to the ideal standards of justification among traditional epistemologists. In spite of this, such "lesser" justifications prompt some degree of acceptance in actual scientific practice. This gradual acceptance of commitments is partly responsible for the differential contribution of commitments to a program (see above).

Given that acceptance of commitments is a matter of degree, the coherentism I have in mind comes with a nuance: not all commitments contribute equally to the coherence of an account or program. To use a metaphor attribute to Sosa (1980), not all planks in a raft contribute equally to its buoyancy. Still, planks need to relate to each other in a configuration that affords overall buoyancy. In the context of commitments, programs, and accounts, this means that reflective equilibrium is not equitable in its assessment of the mutual impact of commitments. Rather, reflective equilibrium is biased towards those commitments with a higher status within a program, i.e., core commitments.

These are not novel ideas. Goldberg (2009) argues that Quinean holism already incorporates this nuance, which manifests as two distinct structures. First, there is a coarse

conjunctive structure, which forms a coherent whole. Second, there is a fine-grained structure, which amounts to the asymmetrical relations of derivation – in particular, inferential derivation – among commitments in the whole.[33] In extrapolating Quine's ideas to my account, his fine-grained asymmetric structure amounts to what I have discussed above as the differential contribution of commitments. These ideas are also often discussed in terms of differential entrenchment (cf. "differential entrenchment of predicates" in Goodman, 1955). Accordingly, Goldberg refers to Quine's approach to coherentism as "holism-cum-entrenchment" (261). To retain consistency with my discussion above, I prefer to discuss this topic as "coherentism-cum-entrenchment."

To show how coherentism-cum-entrenchment shapes the acceptance of commitments, let us consider an example discussed by Quine (1986). The example focuses on the asymmetry of the relations between biological and mathematical beliefs (ibid: 7). Roughly, Quine argues that if a particular biological prediction fails, then the bundle of theoretical beliefs employed in delivering that prediction should be modified. Quine's holism poses that, in principle, any belief in this bundle can be modified, as long as adequate accommodations are carried out elsewhere to retain overall coherence across the bundle. However, the modification of some beliefs demands more accommodations than others, given their larger impact across biology and science at large. In particular, Quine acknowledges the larger impact that mathematical beliefs have across scientific disciplines. Accordingly, he dissuades biologists from modifying mathematical beliefs in order to fix a failed biological prediction. The modification of mathematical principles – say in logic or arithmetic – would reverberate excessively across the whole of science. In contrast, the modification of biological theories has only limited effects in some branches of biology. Thus, when modifying their constellation of commitments, scientists must consider not only the coherence among them, but also their degree of entrenchment. While the previous illustration is set in the context of disciplines and science at large, the same principle can be applied to specific accounts and research programs.

[33] Quine's holism is not explicitly discussed in terms of scientific commitments. However, he discusses entities which qualify as commitments in my approach, namely theories, auxiliary hypotheses, and evidence.

An important theoretical issue of coherentism-cum-entrenchment is the justification of highly entrenched commitments, such as mathematical beliefs in the previous example. There are two sides to this issue. First, there is the coherentist side: highly entrenched commitments, as any commitment, are justified by fitting into a coherent whole. Mathematical beliefs are no exception: they are justified by their fitting into the overall constellation of commitments. Notably, this happens for mathematical beliefs not only in one specific research program or scientific discipline, but rather across the sciences. Furthermore, because of their being highly entrenched, it would be too costly to reject or modify such commitments. That is, several modifications across research programs and scientific disciplines are required to accommodate changes in highly entrenched commitments.

Second, there is the entrenchment side. The questions is: "How are highly entrenched commitments justified in their being entrenched?" In order to answer this question, I introduce a distinction between two notions of entrenchment. On the one hand, entrenched commitments are those that prompt the endorsement of scientists. In other words, the term 'entrenched' tells scientists what to do. On the other hand, entrenched commitments are those that scientists have endorsed systematically. In other words, the term 'entrenched' tells scientists what they have been doing. Accordingly, I refer to these notions as the "normative" and "descriptive" notions of entrenched commitments, respectively.

I suggest that these two notions of entrenched commitments call for different forms of justification. In relation to the normative notion, the acceptance of entrenched commitments is justified by the force of habit or custom (cf. Goodman, 1955). Simply put: entrenched commitments are accepted because they are entrenched.[34] It is worth noticing the dogmatic overtones of this form of justification. In relation to the descriptive notion, the acceptance of commitments that become entrenched is justified by the sustained satisfaction of scientists in using such commitments. Note the practical overtones of this form of justification. Thus, I suggest that highly entrenched commitments admit two forms of justification: practical in their original acceptance, dogmatic in their endurance.

[34] I am expanding the repercussions of entrenchment as originally discussed by Goodman (1955). In his text, Goodman argues that entrenchment of predicates and inferences decides their projectibility. Analogously, in my argument, I suggest that entrenchment of commitments decides their acceptance.

I suggest that these forms of justification – practical and dogmatic – also apply for the justification of core commitments. This claim is not trivial, given that highly entrenched commitments and core commitments are not synonymous. Core commitments are relative to a specific research program. Highly entrenched commitments can cut across research programs and even apply to science as a whole. As a clarification, some highly entrenched commitments – say, an arithmetic theorem – can be core commitments in a research program, e.g., a research program in mathematics exploring the theorem. And, core commitments are – in a sense – entrenched commitments relative to other, more peripheral, commitments within a program.

The proposal that core commitments respond to practical and dogmatic justifications is motivated by the following question: Why should a scientist accept one program as opposed to another? Lakatos provides part of the answer, in his account of why scientists should prefer progressive programs to degenerating ones (I adapt his views to my account). Core commitments are justified in their being retained as core commitments because of a methodological decree. This is equivalent to what I discuss above as dogmatic justification, which is captured by Lakatos's negative heuristics: core commitments are protected from refutation by methodological fiat.[35]

The dogmatic attitude to the justification of core commitments is, in any case, a tamed one: it can be defeated. This is the case when a research program is not able to deliver novel predictions. Or it delivers them, but it is not able to confirm them. In other words, the dogmatic attitude should be abandoned when a program is degenerating. In this case, scientists should move from the degenerating program to a progressive one. In taking this step, there is something implicitly assumed: predictions and confirmation of those predictions

[35] It is debatable whether such form of dogmatic justification is a form of justification at all. For instance, when a parent justifies her decisions to her child by saying "because I say so," it is not clear whether she is providing a justification or withholding it. In this sense, one might construe core commitments as lacking the need of justification in light of recalcitrant evidence: they are safe by decree. Regardless of whether we call it justification or not, the point is that there is a dogmatic attitude that – at least, partly – accounts for the retaining of core commitments.

are highly valued aims across the natural sciences. In this sense, the move from degenerating programs to progressive ones is justified practically.[36]

Beyond the practical justification of moving from degenerating to progressive programs, there is another form of practical justification required to vindicate the choice of a particular program. After all, there might be a plurality of progressive, internally coherent, research programs at any given moment relevant to a scientist, which underdetermines the final decision of a scientist. In this case, I suggest that the aims of a scientist or scientists play a crucial role in justifying their decisions for affiliating to one program or the other. That is, scientists' choice for a program is justified practically.

Taking stock, justification of commitments via coherentism-cum-entrenchment amounts to the following aspects:

i) A single justificatory task can be epistemic or practical. That is, the reasons that provide justification may be bits of knowledge or motivations, respectively.
ii) A justificatory process consists in fitting commitments within a constellation of coherent commitments (in practice, an account or research program).
iii) This justification comes in degrees, given that a commitment can be more or less coherent with an account or program.
iv) Commitments contribute differentially to the coherence of a research program. In particular, core commitments have a starker impact in shaping the process of reflective equilibrium and the resulting coherence.
v) The justification of core commitments involves: i) coherence with other commitments; ii) practical reasons in their being chosen (i.e., motivations); and iii) dogmatic reasons in their being retained (as a heuristic).

[36] There is a sense in which dogmatic justification is only one type of practical justification: dogmatic justification is a good heuristic to engage in progressive research programs (i.e., programs that deliver confirmed predictions). However, I intend to use practical justification in a narrower sense. Scientists' practical motivations tend to be more specific than just adhering to some progressive research program.

To test the merits of the coherence-cum-entrenchment account, I consider the case of the OFC model and the SOC program. As a research program, SOC has gone through dissimilar configurations. These different configurations are the result of modifying, abandoning, and adopting commitments. These changes are informed by the results delivered in new studies from progressively distinct scientific disciplines. These changes, I suggest, are a manifestation of reflective equilibrium in play, which increments the overall coherence across the program.

This can be illustrated in further detail by focusing on four theoretical commitments that were originally part of the SOC program (for more details, see Chapter 2): i) [C1] is that spatial and temporal fractals are unavoidably related; ii) [C2] is that dissipative systems organize themselves towards a critical state in which power-law correlations occur; iii) [C3] is that a system must be open, extended, dissipative, and slowly driven to exhibit SOC; and iv) [C4] is that spacetime fractals are snapshots (i.e., a manifestation) of SOC.

Watkins *et al.* (2016) discuss how these commitments have evolved into their present configuration: i) C1 was abandoned because it has been proven false; ii) C2 involved a new and idiosyncratic technical sense in which the term 'critical' was being used in the context of SOC; iii) C3 has been retained but with further refinements due to new studies; iv) C4 has been proven partially wrong and, accordingly, its scope has been refined. This evolution has been shaped by the mutual impact that cumulative experiences and results in the context of the SOC program have had, i.e., via reflective equilibrium.

It is tempting to say that early practitioners in the SOC program were epistemically justified in accepting C1 to C4. That is, these theoretical commitments were justified *to the best of their knowledge*. However, a closer look reveals some nuances. For example, consider C4. In an early paper in the history of the SOC program, Bak and Chen (1989) state that spacetime fractals are snapshots of SOC (5). To justify this claim, Bak and Chen present and discuss various simulations with cellular automata whose results can be described as fractals. These models and their results are supposed to provide epistemic justification to Bak and Chen's claims. However, the fractal behavior of these models is produced in the presence of spatial correlations, expressed in the models as avalanches. Discovering a mechanism that

produces fractal behavior in a cellular automaton is not equivalent to discovering the physics of all spacetime fractals in nature. In other words, Bak and Chen have not showed that all spacetime fractals in nature are produced via SOC, only some of them do. In this sense, Bak and Chen's commitment to C4 is not epistemically justified.

Instead, I suggest that C4 is justified practically: it is a convenient hypothesis to guide the efforts within the SOC program. It is convenient because it captures the generalist or unificationist motivations of the scientists interested in discovering the physics of fractals. The epitome of this attitude is reflected in the title of Bak's 1996 "How Nature Works." And it did work. C4 attracted the attention of scientists beyond physics. This does not mean that C4 was without its problems: it produced – and still produces – a significant amount of confusion. Watkins *et al.* (2016) devote a whole section of their paper to the topic (Section 8.1). In a personal communication, researcher Gunnar Pruessner – author of "Self-Organized Criticality: Theory, Models and Characterization" (2012) – commented on the issue:

> *I think Per Bak did not do himself a big favour by being so outspoken, so fierce in his promotion of his own subject. I think he came across as a really aggressive proponent of SOC. On the other hand,* that got a lot people of people really going. *I think scientists […] were wind up by people like Per Bak who say "Look, it's all SOC". His book "How Nature Works" … a friend of mine used to say: "Well, it sounds like it is the first volume in a whole series and the next volume is 'How Cars Work'".*

One reason why the modifications of C1 to C4 have been rather excruciating is their status as original core commitments of the SOC program. They are what Watkins *et al.* (2016) refer to as the "SOC hypothesis." Given their central position in the program, the assessment of the overall coherence of the program was heavily biased towards them. Because of this, I suggest that these commitments were protected from refutation by methodological fiat (i.e., dogmatic justification). In practice, this means that significant evidence – either in size, rigor, or scope – had to be accepted within the program in order to modify or remove these core commitments. Still after three decades of research, much controversy and confusion surround these commitments. Watkins *et al.* (2016) have provided some guidance on what should be retained from the original SOC hypothesis for the next stages of the program.

The reconfiguration of the SOC program can also be explored in light of the contributions of the OFC model. I focus on one particular contribution, namely the role of conservation in SOC. Before OFC's paper, most SOC simulations with cellular automata assumed a conservative regime, i.e., transference during avalanches conserves quantities. In fact, it had been suggested that conservation was a necessary condition for SOC (e.g., Hwa & Kardar, 1989; Grinstein *et al.*, 1990; Dhar, 1990; Manna *et al.*, 1990). I refer to this as the "conservation hypothesis." OFC falsified this hypothesis: they presented a nonconservative model that exhibited SOC. (In fact, this was suggested before by OFC's precursor: Feder and Feder, 1991.)

Hence, it is worth reflecting on why the OFC model did not take long to be accepted as an exemplar of SOC research. I identify two main reasons for this. First, OFC's contributions to the SOC program did not challenge its core commitments. Quite the opposite: OFC acknowledge the influence of the SOC program in the design of their model. The OFC model was built upon core ideas of the program, but it extended them and explored different simulation conditions. In particular, I suggest that the conservation hypothesis was not a core commitment in SOC at the time in which OFC introduced their model. After all, the conservation hypothesis was not explicitly stated in BTW's two seminal works. Watkins *et al.* (2016) also do not recognize the conservation hypothesis as part of the original SOC hypothesis. In this sense, I suggest that the conservation hypothesis should be described as a peripheral line of research within SOC, a line that had to come to terms with OFC's results. In the end, OFC's falsification of the conservation hypothesis was convincing enough to dismiss the hypothesis.

This brings me to the second reason why the OFC model – and its contributions – did not take long to be accepted, even though they falsified the conservation hypothesis: the quality of OFC's research was compelling. In particular, the fact that it was based on a well-reputed model – namely the BK model – helped establishing its own reputation. This is noted by Pruessner (2012): "Within statistical mechanics, the OFC Model is a unique SOC model, incorporating, to the great surprise of many, non-conservation and all features that are quintessential ingredients for SOC models: slow drive and avalanching (separation of time

scales), thresholds, local dynamics. The *credibility* of the OFC Model benefits greatly from its geophysical origin as the Burridge-Knopoff Model. To date, earthquakes are a showcase natural phenomenon for SOC" (139; my emphasis). The fact that OFC model cohered with well-established lines of research in seismology allows the SOC program to extend its scope coherently to other disciplines. In other words, OFC respond to the generalist and unificationist attitude discussed above. They provide results that cohere with lines of research in other disciplines *and* with the results of the SOC program, thus extending the scope of the program.

Thus, there are two considerations in the acceptance of OFC's new contributions as SOC commitments: i) the overall coherence of OFC's contributions with the SOC program, together with the peripheral position of the conservation hypothesis, which was incoherent with OFC's nonconservative principle; and ii) the merits of OFC's nonconservative principle, together with its coherence with well-reputed lines of research in other disciplines. I suggest that these considerations are expressions of the coherence-cum-entrenchment scheme of justification and acceptance of commitments.

With this, I finish this chapter. In the following one, I explore a specific set of scientific commitments, namely explanatory commitments. They play a central role in my argument, given my focus on explanation as a scientific enterprise.

3.4 Précis

The main theoretical theses of this chapter are:

- Thesis 1: Scientific commitments are general guidelines, accepted by working scientists, which shape their research.
- Thesis 2: Scientific commitments are organized in (meta)stable bundles known as "accounts." These accounts are adopted as part of collaborative endeavors of scientific research known as "research programs." Commitments have a differential contribution to scientific research. This translates into the existence of core and peripheral commitments.

- Thesis 3: In programmatic contexts, scientific commitments are justified epistemically or practically, in accordance to a coherence-cum-entrenchment scheme.

The main lessons learnt from my case studies are:

- BK express commitments in their research. For example, they are committed to i) theoretical and ii) model methodologies.
- BK's commitment exhibit a differential contribution to their research. In particular, distinct modeling principles are more or less peripheral to their research.
- Core commitments within the SOC program have evolved in accordance with processes of reflective equilibrium. The original acceptance of some core commitments – e.g., C4 – responds to practical motivations. The core of the program has been resilient to change due to dogmatic justification. New commitments – such as OFC's nonconservative principle – are justified in light of a coherence-cum-entrenchment scheme.

***

4 Explanatory Commitments and Accounts of Explanation

In this chapter, I introduce and discuss a specific kind of scientific commitment, namely "explanatory" commitments. Explanatory commitments are commitments in exactly the same sense introduced in the previous chapter: They are general guidelines, accepted by scientists, which shape their research. What is specific about explanatory commitments is that they are accepted to guide the crafting of explanations and, through them, related epistemic achievements, such as prediction or understanding. The notion of explanatory commitments is intended to capture a ubiquitous observation: scientists across different research programs are typically committed to different standards about what constitutes a satisfactory explanation. With this, I do not only mean those standards that serve to judge the adequacy of the content of an explanation. Rather, I mean those standards that serve to evaluate the way in which the content is presented or put together to function as an explanation.

The structure of this chapter is the following. In section 4.1, I address two general questions concerning explanation, namely what they are and what makes them good. I frame my discussion in terms of a well-established debate between ontic and epistemic conceptions of explanation. I study how these conceptions are applied in the context of the SOC program and, in particular, in the OFC model. In section 4.2, I address the topic of accounts of explanation. I introduce the notion of accounts of explanation, provide some examples, and establish the relation between them and explanatory commitments. I discuss the latter issue in the context of the BK model. In section 4.3, I discuss some of the problems that emerge from using accounts of explanations as tools for description and evaluation of explanatory practices. I suggest that more accurate descriptions and nuanced evaluations of explanatory practices can be made by adopting a more fine-grained approach which focuses on explanatory commitments. Eventually, these explanatory commitments can be classified in terms of well-established accounts of explanation, at the cost of missing some of the nuances of the explanatory practice in the process. I illustrate these ideas in the context of my case studies.

4.1 The Ontic and Epistemic Conceptions of Explanation

There are two main approaches to what explanations are, namely the epistemic and ontic views of explanation. On the one hand, the epistemic view posits explanations as epistemic activities that increase our understanding and knowledge of a phenomenon (Wright, 2012: 382). These epistemic activities are embodied in acts of communication, in the reading of texts and models, and by means of entertaining mental representations (cf. Craver, 2014). From an epistemic point of a view, explanations are subjective (or, at the very least, intersubjective). This implies that a phenomenon might have many distinct explanations, depending on the different texts, models, or mental representations that we use to understand it. On the other hand, the ontic view of explanation holds that explanations are not a text, model, or mental representation, but an objective part of the causal structure of the world, a "full-bodied" thing (Craver, 2007: 27). From an ontic point of view, phenomena in the world have explanations, even if we do not know them. The role of science is to discover these explanations.

Despite the obvious tensions between these views, there are ways in which they actually complement each other. On the one hand, the epistemic view arguably relies on the existence of a real causal structure to be described and communicated (that is, if one holds a realist stance). In this sense, the epistemic view often presumes the ontic view. On the other hand, advocates of the ontic view may well aim for the discovery, description and communication of the causal structures that explain a particular target phenomenon. In this sense, the ontic view is compatible with the goals of the epistemic view.

In recent years, the emphasis of the debate has changed from discussing the nature of explanations to the conditions that make an explanatory text a successful explanation. In other words, the debate has shifted from a demarcation problem to a normative one. From an ontic point of view, explanatory texts are successful explanations if and only if they describe the causal structure of the world (Illari, 2013: 250). From an epistemic point of view, explanatory texts are successful if and only if they advance our understanding and knowledge of a phenomenon of interest (ibid). In sum, instead of discussing which view captures better

the nature of explanation (i.e., demarcation problem), the debate nowadays focuses on which normative constraints, ontic or epistemic, are pre-eminent (i.e., normative problem).

Illari (2013) attempts to overcome the ontic-epistemic dichotomy by showing that advocates of the ontic and epistemic views actually hold both types of considerations in assessing the success of explanations. Illari reviews works by Craver (an advocate of the ontic view) and Bechtel (an advocate of the epistemic view) and reveals that both of them accept "rival" constraints in their accounts of successful explanations. Illari concludes that no single view on the nature of explanation has normative priority over the other. Van Eck (2015) describes Illari's thesis as follows: "[…] good explanations are subject to both ontic and epistemic constraints: they must describe mechanisms in the world (ontic aim) in such fashion that they provide understanding of their workings (epistemic aim)" (5). Accordingly, Illari's thesis is often referred to as the "integration" account (van Eck, 2015).

The integration of ontic and epistemic views has not been universally accepted. Although ontic and epistemic considerations seem to be part of good explanations, there is still debate on issues of emphasis or priority. For example, van Eck (2015) argues that the aims of the ontic and epistemic views rely on an epistemic prerequisite, namely the discovery of causal roles functions. And Sheredos (2016) argues that the relation between the ontic and epistemic views is not one of integration, because they are autonomous. Furthermore, in some explanations epistemic considerations are prior (e.g., in general explanations), and in other explanations ontic considerations are prior (e.g., single-case explanations).

For the purposes of my argument, I do not need to take a stance in the demarcation side of this debate. That is, I do not need to establish whether the nature of explanations is ontic or epistemic. With regards to the normative side – i.e., what makes for a good explanation – I assume a pragmatic and contextual approach. My views follow from my previous discussions on scientific commitments in research programs. In this view, what makes for a good explanation is relative to the commitments held by scientists within a program. This includes scientists' commitment to an ontic or epistemic approach to explanation. More explicitly, if scientists in a research program are committed to the ontic view, then a good explanation is that which captures the actual causal structure of the

explanandum phenomenon (whatever that means in the context of the program). And, if scientists in a research program are committed to the epistemic view, then a good explanation is that which affords understanding via a cognitively expedient representation.

I explore these ideas in the context of the SOC program and its explanatory agenda. To begin with, it is important to establish that explanation is a central epistemic aim of the SOC program. From its outset, SOC was intended as an explanatory concept. This is evident in the very title of BTW's 1987 seminal paper, which posits SOC as "an explanation of 1/f noise." In this paper, BTW highlight the lack of explanations of two – presumably related – phenomena, namely 1/f noise and fractal structures: i) "Despite much effort, there is no general theory that *explains* the widespread occurrence of 1/f noise" (Bak *et al.* 1987: 381; my emphasis); ii) "Another puzzle seeking a physical *explanation* is the empirical observation that spatially extended objects, including cosmic strings, mountain landscapes, and coastal lines, appear to be self-similar fractal structures" (Bak *et al.* 1987: 381; my emphasis). The concept of SOC is supposed to close this explanatory gap: "In 1987 [BTW] developed a concept to *explain* the behavior of composite systems, those containing millions and millions of elements that interact over a short range. [BTW] proposed the theory of self-organized criticality: many composite systems naturally evolve to a critical state in which a minor event starts a chain reaction that can affect any number of elements in the system" (Bak & Chen, 1991: 46; my emphasis).

SOC has also been posed as a potential explanation of more specific natural phenomena and their features. A notable case should be familiar by now: the occurrence of earthquakes. This is illustrated in the following quotes: i) "Self-organized criticality may *explain* the dynamics of earthquakes, economic markets and ecosystems" (Bak & Chen, 1991: 46; my emphasis); ii) "The theory of self-organized criticality has been successful not only in *explaining* the evolution of earthquakes but also in describing the distribution of the epicentres of earthquakes" (Bak & Chen, 1991: 51; my emphasis); iii) and "Bak and Tang indicated that the simple conservative SOC models can serve as a framework for *explaining* the power-law behavior [of earthquakes]" (OFC: 1244; my emphasis).

In particular, OFC also consider their model as providing explanations of aspects of earthquake occurrence, namely the power-law behavior and the variability of exponents in those power-laws: "[…] our results, apart from providing an *explanation* for the observed power laws, also give some *explanation* for the observed variability [of B-values]" (Christensen & Olami, 1992a: 1835; my emphasis). This is asserted more explicitly in the following quote: "the dependence of the power laws on the conservation allows us to *explain* the wide variances in the Gutenberg-Richter law as a result of the variances of the elastic parameters" (1244; my emphasis). OFC not only claim that their model explains aspects of the behavior of earthquakes. They also claim that it does so better than previous attempts. For example, Christensen and Olami (1992a) submit that the OFC model – by being nonconservative – overcomes the failures of previous attempts: "Most of [the previously] suggested models are conservative and have no physical interpretation in the context of the driven spring-block model. Furthermore, since the models are conservative, they predict unique power-law exponents which are much lower than the observed values (B = 0.10)" (1835).

Having argued that explanation plays a central role in the SOC program – including OFC's research – I proceed to discuss how explanation is conceived within the program. Originally, researchers in the SOC program expressed ontic motivations in their approach to explanation. That is, explanations of SOC phenomena were intended to capture the causal structure of complex systems exhibiting SOC. There are various cues that hint at these motivations. Most notably, SOC was posed as the mechanism responsible for fractal phenomena: "We suggest that this self-organized criticality is the common underlying mechanism behind the phenomena described above [namely 1/f noise and spatial fractal]" (BTW, 1987: 381). The ontic attitude towards explanation is also manifest as the SOC mechanism is said to have been "discovered" (e.g., Bak, 1996: Chapter 2 "The Discovery of Self-Organized Criticality").

Still, this ontic view towards explanation is a tame one. Researchers in the SOC program often admit that their explanations are simplified descriptions of the actual ontic explanation. This is expressed in Bak's discussion of the "philosophy of using simple models" (1996: 41). Or consider the following quote: "The models are not realistic representations of

any real systems. We have chosen to sacrifice realism for simplicity in order to obtain a general flavor of the mechanisms at work" (Bak & Chen, 1989: 6). This preference for simplicity is a decision that aligns with an epistemic approach to explanation: a good explanation is one that affords understanding.

By expressing contentment with a "general flavor" of the actual mechanism, the ontic views of explanation are tamed with a more epistemic attitude. Accordingly, I suggest that SOC's explanatory agenda seems to be better described by an integration of both ontic and epistemic views of explanation. SOC researchers aim to describe an actual mechanism in the world, but to the extent that the explanation affords them with understanding. Thus, they are willing to engage in distortions and simplifications. This integration confirms Illari's thesis.

Still, this does not mean that both components have been equally well received by the scientific community. In fact, the explanatory achievements of the SOC program have been a matter of great controversy. I suggest that most of these controversies are due to the ontic component of SOC's explanatory attitude. In particular, there is wide criticism towards the construal of the central explanandum phenomenon, namely ubiquitous fractals in nature. A main problem is that this explanandum seems to be wrongly construed as ubiquitous.[37]

One of the few philosophers of science who has tackled this issue is Frigg (2003).[38] He reports a series of examples of a "generalist" attitude within the SOC program (614). Such optimistic attitude concerning the scope and explanatory power of SOC is – according to Frigg – based on a misguided appreciation of the ubiquity of SOC behavior. Frigg claims that the alleged ubiquity of SOC amounts to the recognition that a core set of models, related by

[37] Along these lines, Bokulich (2018) argues that the ontic conception of explanation is fundamentally misguided. This is because what is explained is not an objective part of the causal structure of the world. Rather, what is explained is a particular conceptualization of an explanandum phenomenon in the context of a research program. As a result, she advocates a particular form of the epistemic conception of explanation, known as the "eikonic" conception.

[38] Frigg (2003) highlights the total absence of philosophical publications discussing SOC phenomena until his paper (614-5). While his paper introduced the topic to philosophers of science, it did not initiate a significant movement of philosophical discussions and contributions regarding SOC. His paper has been cited by 50 research articles, with only three of them published in philosophy-oriented journals (up to April 2018). In particular, to the extent of my research, I have not found philosophical contributions on explanations of SOC. My case studies are intended to close this gap.

formal analogies, are used to represent a wide variety of phenomena. In addition, these models are able to simulate such phenomena with statistical relevance. Although Frigg admits the representational versatility of SOC models, he argues that most of these models represent their targets in extremely idealized fashion. Furthermore, the way in which SOC models simulate natural SOC phenomena is not always the actual way in which natural SOC phenomena unfold. That is, SOC models – for the most part – only provide a possible mechanism for generating the observed SOC phenomena in nature. Hence, Frigg's conclusion is that the ubiquity of SOC is mistakenly inferred from the vast representational capacity of a set of SOC models.

Watkins *et al.* (2016) deliver similar criticism. They argue that the alleged ubiquity of SOC is not even based on vast observational or experimental data:

> *"Where fractals and scaling are suspected in natural phenomena, observational support is often very limited (Avnir* et al. *1998, e.g.), both in terms of length and time scales spanned by the data as well as its robustness. Broad distributions are frequently found, but there are few phenomena, which offer sufficiently detailed and broad data to support power law scaling beyond reasonable doubt. It is difficult to reconcile the efforts that have been spent on experiments, data gathering and analysis with the claim that scaling or just power laws are ubiquitous in nature. One may therefore ask, rather provocatively: Is there really a (ubiquitous) problem to solve?" (26).*

In addition, Watkins *et al.* (2016) submit that the core claim of the SOC program – namely that self-tuned phase transitions exist in nature – has received computational confirmation, but no unambiguous or unquestioned evidence for it (9). However, they add that "it is also fair to say that experimental, observational, numerical and analytical work is homing in to corroborate at least the core claim" (9).

There is a myriad of SOC models whose names contribute to the misconception of the ubiquity of SOC. For example, "sand pile" models, "forest fires" models, and "evolution" models, to name a few. Researcher Gunnar Pruessner adopts a pragmatic attitude towards these names: they play a mnemotechnic function (personal communication). Consider the

"forest fire" SOC model as an example. It is often claimed that forest fire SOC models are models that represent forest fires (e.g., see Frigg, 2003). This is a reasonable assertion, although one that is not completely accurate. The forest fire model is actually intended as a model of general turbulence. It is often presented in terms of forest fires for pedagogical effects: it can be explicated in terms of trees being lightened up in chain reactions. In this sense, the forest fire model of turbulence is a model of forest fires in the same way that – for example – the "plum pudding" model of the atom is a model of plum puddings. The label "forest fire" plays the same role as the label "plum pudding": It is a mnemotechnic device.

In light of Frigg (2003) and Watkins *et al.* (2016)'s comments, I argue the following. Broadly speaking, scientists in the SOC program are interested in explanations of spatial and temporal power-law correlations. Their approach to explanation embodies ontic motivations: their explanations are intended to capture the actual causal structure of SOC phenomena. However, SOC researchers do not – for the most part – deliver actual explanations of SOC phenomena. And they seem to be aware of this, up to a certain point. There are three senses in which the ontic motivations fail to obtain. First, the very explanandum phenomenon is non-actual, but an idealized construal (this is Watkins *et al.*'s point). Second, even in exemplary cases of target SOC phenomena, the constructed models only approximately simulate some features of the target. In this sense, SOC models are – in the most promising scenarios – approximate and partial explanations. Third, some SOC models produce their outcomes in ways that depart from the actual processes in the target phenomena (this is Frigg's point). This is the case even in situations where the simulated outcomes statistically resemble the empirical data. In this sense, SOC models should be thought of as providing only possible or maybe plausible explanations.

Taking into account the early ontic motivations in the SOC program, explanations of SOC behavior based on SOC models are simply bad explanations. This is because they do not capture the actual causal structure of SOC phenomena, for the various reasons discussed above. To be sure, the limitations of the ontic attitude have been acknowledged by most researchers in the field, who progressively have adopted a more epistemically oriented attitude. From an epistemic view, explanations delivered in the SOC program are legitimate:

they provide representations that are cognitively accessible and advance our understanding of complex systems exhibiting SOC.

4.2 Accounts of Explanation and their Relation to Explanatory Commitments

Another way in which philosophical debates on scientific explanation have organized is around the notion of "account of explanation," also referred to as "theories" or "models" of explanation (e.g., see Woodward, 2019). Accounts of explanation can be characterized as different explications that philosophers provide of the concept of 'explanation' (Weber *et al.*, 2013: 25). That is, the various accounts of explanation highlight distinct aspects of the concept of 'explanation', construing it somehow differently but in a precise manner. From a more empirical approach, accounts of explanation may also be characterized as descriptions that philosophers provide of explanatory practices. In fact, it has become customary in the philosophical literature to discuss the merits of accounts of explanation in light of the evidence provided by case studies.

Some philosophers of science have attempted to construct synoptic narratives about the evolution of the debates on accounts of explanation. One of the most authoritative overviews is Salmon's "Four Decades of Scientific Explanation" (1989), which remains relevant to this day. For more recent overviews, see Weber *et al.* (2013) or Woodward (2019). Roughly, Salmon's narrative is one of a shift from an old consensus, founded on the deductive-nomological (DN) account of explanation (due to Hempel and Oppenheimer), to a state of vast dissensus. Some of the most eminent accounts that are part of Salmon's state of dissensus are the causal-mechanical account (Salmon, 1998; Dowe, 1992), the unificationist account (Friedman, 1974; Kitcher, 1989), and the pragmatic account (van Fraassen, 1980; Achinstein, 1983). Since the publication of Salmon's essay almost three decades ago, new and distinct accounts of scientific explanation have been proposed. Some examples are the manipulationist account (Woodward, 2003), kairetic account (Strevens, 2008) and the new mechanical account (Bechtel & Richardson, 2010; Glennan, 1996; Machamer, Darden & Craver, 2000), among others. Hence, the dissensus – or, rather, the plurality – remains.

I do not attempt to review these various accounts of explanation. However, I consider it important – for the purposes of my argument – to point at some of the aspects in which these accounts differ from each other. To begin with, different accounts of explanation take explanations to be different things. For example, explanations can be arguments (e.g., Hempel's DN account; 1965), descriptions (e.g., new mechanist account, particularly Glennan's approach; 2017), processes (e.g., Mantzavinos's explanatory games account; 2016), answers to why-questions (e.g., van Fraassen's pragmatic account; 1980), just to name a few cases. This plurality reflects the various kinds of things that we take to be explanatory. As Faye (2014) suggests, we might even take people, facts, events, hypotheses, models, and theories to have explanatory power (114).

Each account of explanation construes "explanation" as having a certain ontology, structure, and inner logic. For example, the DN account has as component elements "premises," "laws of nature," and a "conclusion," which are related in an argument by entailment. As another example, the elements of a mechanistic explanation include entities, activities, organization, and phenomenon, and they are related causally and constitutively in a well-structured description. A final example may be given in the context of the pragmatic account of explanation. In this case, the constitutive elements are a question (composed of a topic, a contrast class, and a relevance relation) and an answer, which stands in some relevance relation to the question.

Some philosophers argue that there is a basic common logic to all accounts of explanation, namely the dependence relation between explanandum and explanans. This relation has been cashed out in different ways (e.g., as the dependence relation between problem and solution, cause and effect, or question and answer). However, even this seemingly basic aspect of the logic of explanation has been debated. This is exactly what pragmatic approaches to explanation do. Most pragmatist views argue that explanation is not an objective relation between explanandum and explanans. Rather, the role of the user, her background knowledge, and her intentions are inextricable from an adequate characterization of explanation. In this sense, the logic of explanation is not a two-place structure, but rather a three-place one. This example serves to emphasize the diversity of approaches to the notion of explanation, even concerning its most abstract features.

After having introduced the notion of accounts of explanation and illustrated their variety, I proceed to propose a relation between this notion and my main concern, namely explanatory commitments. In order to do this, I must engage in some terminological clarifications. 'Accounts of explanation', as I have been using the term, are philosophers' attempt to articulate scientists' explanatory practices (call this the "philosophical" sense). However, the term 'accounts of explanation' may also be taken as the very explanatory practices of scientists (call this the "scientific" sense). In the scientific sense, accounts of explanation are accounts in Elgin's sense (2017): accounts of explanations are constellations of commitments standing in mutual support. Thus, the philosophical and scientific senses of accounts of explanation relate to each other as model relates to target.

As a consequence, the plurality and variety of accounts of explanation (in the philosophical sense) reflect two considerations: i) philosophers describe a wide variety of accounts of explanation (in the scientific sense); and ii) philosophers differ in their descriptions of roughly similar accounts of explanation (in the scientific sense). Because of this, I suggest that those well-established accounts of explanation (in the philosophical sense) may be conceived as: i) descriptions of major and stable accounts of explanation (in the scientific sense) or ii) general purpose descriptions which may be used to describe a wide range of accounts of explanation (in the scientific sense) more or less accurately.

My views on this topic resemble – in some respects – those of Mantzavinos (2016) and Weber *et al.* (2013). In Mantzavinos's view, explanatory practices across the sciences are not well captured by any single account of explanation (in the philosophical sense). Instead, explanatory practices amount to a plurality of explanatory games. These games possess their own rules, which are decided by the "players" in the game and they might evolve over time. These rules shape the outcomes of the explanatory practices delivered in the context of a particular game. Mantzavinos's notion of explanatory games is not radically different from that of accounts of explanation (in the scientific sense). But it emphasizes the contextual and pragmatic nature of explanatory practices. I submit that accounts of explanation (in the philosophical sense) should be regarded as attempts to characterize the "rules" that regulate

explanatory practices in different contexts (Mantzavinos's explanatory games). However, I prefer to use the broader notion of "explanatory commitments" as opposed to only rules.

In the following, for matters of economy, I use the term 'account of explanation' *simpliciter* to refer to accounts of explanation in the philosophical sense. Whenever I intend to use the same term in the scientific sense, I will make that explicit. I also resort to the term 'explanatory practice' to refer to the manifestation of an account of explanation in the scientific sense.

For their part, Weber *et al.* (2013) propose an "approach" to scientific explanation, which is different from the notion of an "account" of explanation. As part of this approach, they argue that the various accounts of explanation available in the literature should be taken collectively as a "toolbox." This means that the various accounts of explanations may be used as tools to describe and evaluate explanatory practices. Furthermore, Weber *et al.* acknowledge that the employment of accounts of explanation as tools for description and evaluation is contextual. I think Weber *et al.*'s project goes in the right direction. Nonetheless, I have some reservations which I discuss further below.

To test my views on the relations between accounts of explanation and explanatory commitments, I explore these issues in the context of the BK case. My strategy is to identify BK's explanatory commitments and relate them to well-established accounts of explanation. From the outset, I admit that this is a contentious task. There is a significant and arguably irreducible interpretative component in conducting this enterprise. Indeed, BK do not declare explicitly which account of explanation best captures their explanatory practices. To be fair, most scientists would not do this. Describing explanatory practices in terms of accounts of explanation is – for the most part – a philosophers' game. Having said this, I suggest that BK do engage in tasks that express definite explanatory commitments. In this sense, the fact that this section is interpretative does not make its content arbitrary. My interpretation is based on cues – namely terminology and reported practices – that reflect explanatory commitments. As a result of conducting this exploration, I conclude that BK's explanatory commitments can be classified in terms of a well-established account of explanation, namely the new mechanist account.

The first step is to show that BK do have explanatory motivations in constructing their model. I submit that they do, although they are less vocal about this than SOC researchers and OFC in particular. The term 'explanation' is used on only two occasions in the original paper. On one occasion, it is used to refer to the explanatory role of linear theories developed in mathematical physics (341). Subsequently, BK submit that most phenomena related to the occurrence of earthquakes is extremely nonlinear. In this sense, it is plausible to interpret BK's project as filling the explanatory gaps left by linear theories. The other occasion on which BK use the term 'explanation' is to refer to the explanatory role of qualitative theories of the occurrence of earthquakes. Given that the starting point of the BK model is qualitative theories of the occurrence of earthquakes, it also seems plausible to argue that BK's efforts are a continuation of the explanatory agenda of the qualitative theories. Furthermore, some of BK's conclusions seem to point at the explanatory role of their model. Consider the following quote: "[…] if the demonstrations of the laboratory and numerical models are borne out in nature, it would seem likely that the nature of the friction on a fault surface *determines* the statistical properties of the earthquake shocks that are observed" (370; my emphasis). I suggest that the term 'determine' in this quote can be read as playing an explanatory role.

The next step in my argument is to identify BK's explanatory commitments. Cues concerning this issue can be obtained by scrutinizing BK's actual object of study, namely the spring-block model. Indeed, the laboratory and mathematical implementations of the BK model share a central feature: both of them are based on a spring-block system. The laboratory model is a material construction of a concrete spring-block system. And the mathematical model is a physico-mathematical description of a general version of the spring-block system, in tandem with a numerical computer model that solves the physico-mathematical description. Given that BK are interested in studying the occurrence of earthquakes, it is telling that they decide to employ a spring-block system as a surrogate system.

In discussing the laboratory implementation, BK admit that there is a practical motivation in constructing the spring-block system. BK claim that the construction of the spring-block system – as opposed to the imagined continuous elastic string that they discuss

earlier in their paper – is preferred "[f]or mechanical convenience" (344). This is a telling decision, especially considering that a continuous string would have better captured a commonplace assumption in seismology, namely that of treating rocks as a continuum. To be clear, the mechanical convenience to which BK allude is not mere ease in assembling a spring-block system. Rather, the mechanical convenience points at the advantages that such system affords in research.

The spring-block system is particularly advantageous for two purposes, namely analysis and manipulation. Consider the issue of analysis of shocks. Shocks in the spring-block model are individuated each time blocks slide. Observing and keeping track of sliding events in a spring-block system with eight blocks is simpler than observing and keeping track of sliding events in a continuous string with indefinitely many points. In addition, the size of a shock (i.e., its energy release) is a function of the change in position of blocks. To calculate the energy release in a continuous string, again, would involve calculating energy release after slide events for indefinitely many points. Now, consider the issue of manipulation of the model. For example, take into account the springs' constants. In the first laboratory simulation, all springs' constants are the same. In the second laboratory simulation, the springs' constants are graduated: the spring with the smallest constant is closer to the motor. This setting is established by literally cutting the springs to adjust their constants, i.e., via manipulation. An analog intervention in a continuous elastic string would be rather challenging.

There are two features of the spring-block system that are particularly relevant in affording expedient analysis and manipulation, namely its discreteness and modularity. Given that the spring-block system is discrete, observations can be efficiently focused on a finite number of entities. This facilitates analysis of events that occur during the simulation. And, given that the spring-block model is modular, BK can intervene the model and test different elastic coefficients and laws of friction for each block.

I suggest that these two features of the spring-block system – discreteness and modularity – are a consequence of BK's acceptance of two guiding principles, namely P1 and P2 (see Section 3.2). P1 states that a discrete system is a suitable surrogate to simulate the

behavior of a continuous medium. And P2 states that friction is a causal factor that suffices to account for the statistics of earthquake behavior. P1 and P2 are principles accepted by BK, that are intended to be tested and eventually demonstrated in their model. Accordingly, the design of the BK model is guided by P1 and P2.

Consider P1. This principle is tested differently in the distinct implementations of the BK model, but the overall logic behind the demonstrations is analogous. The idea is to construct (in the laboratory implementation) or describe (in the mathematical implementation) a discrete dynamical system. Simulations are run in the constructed or described system, respectively. The results of the simulations are then compared to empirical statistical data of real earthquakes. If the simulations' results resemble the behavior of real earthquakes – under some agreed standard – then the principle is demonstrated. Thus, it is commitment to P1 which prompts the discreteness of the BK model.

Now, consider P2. In order to demonstrate this principle, BK isolate friction as a causal factor for a focused investigation. Friction is represented in the spring-block system as a force between the blocks and a rough surface. In this sense, frictional forces are localized on the blocks: blocks are the entities exerting the causal role of resistance to movement as they interact with the rough surface. This is especially clear in the mathematical implementation of the BK model: there is a frictional force assigned to each block in the equations of motion. Furthermore, because friction is localized on the blocks, different laws of friction can be assigned to each block independently. This feature is exploited by BK in the context of their discussion on aftershocks: BK define three distinct laws of friction to be applied to three different sections (i.e., sets of blocks) of the spring-block system (see Figure 2-9). In this sense, commitment to P2 prompts the modularity of the BK model.

I suggest that acceptance of P1 and P2 prompts the acceptance of discreteness and modularity as explanatory commitments. This suggestion emerges from combining the two previous results: i) the BK model is used for explanatory purposes; and ii) the BK model is designed as discrete and modular (in response to P1 and P2). To put it more bluntly, BK's explanations derived from their model rely on a discrete and modular construal of a geological fault.

My intention now is to relate these explanatory commitments to a well-established account of explanation. My suggestion is that discreteness and modularity may adequately be ascribed to a new mechanist account. I will not go into a detailed review of the new mechanist account nor the internal debates. Still, there are few platitudes on which most new mechanists would agree. In the context of the new mechanist account, explaining a phenomenon amounts to describing the mechanism responsible for it. Roughly, a mechanism is a set of entities and activities that, organized in a certain way, are responsible for a phenomenon (e.g., Illari & Williamson, 2012: 123).

My classification of BK's explanatory commitments as mechanistic is originally prompted by BK's claim concerning the *mechanical* convenience of the spring-block model. However, beyond this marginal remark, there is a more important consideration: the design of the BK model as a discrete and modular system aligns with a new mechanistic methodology of mechanism description. The methodology in question is known as "decomposition-localization" (Bechtel & Richardson, 2010). The core of this method can be roughly explained as follows. In order to study a complex system, one must break the system down into parts (i.e., decomposition) and assign causal roles to them (i.e., localizing). It seems accurate to say that this is exactly what BK do with their model: they identify the parts of the spring-block system and assign causal roles to them. This is particularly evident in BK's mathematical description of the spring-block system. BK decompose the system in blocks: each block has its own equation of motion. And the various activities are allocated to each block. Furthermore, different activities can be attributed to different blocks. This feature is actually exploited as BK test distinct laws of friction in distinct blocks. In addition, decomposition and localization also covers the individuation of springs, which are allocated with a particular elastic coefficient. And a driving force is allocated to the motor (in the laboratory implementation) or moving slab (in the mathematical implementation).

Furthermore, the spring-block model seems to be adequately characterized as a mechanism, in the new mechanist sense. There are entities, such as the blocks, the springs, the rough surface, and the motor (or pulling slab in the mathematical implementation). There are activities, such as the slow pulling of the motor (slab in the mathematical

implementation), the deformation of springs, the sliding of blocks and the consequent pulling or pushing of their neighbors. And there is an organization, manifested in the arrangement of the blocks and springs in a one-dimensional lineal array, its location on top of a frictional surface, the pulling on one side and a free end on the other side, and the neighborhood organization in which each block has two neighbors, with the exception of the first and last block in the system. The main explanandum phenomenon is the various shocks produced by the spring-block system and the statistical patterns that these shocks embody. For these reasons, I suggest that BK's explanatory commitments can be safely classified as mechanistic.

A sceptic might criticize my classification of BK's explanatory commitments as mechanistic. She might argue: "you have only shown that BK engage in mechanistic modeling but have not shown that BK intend to produce mechanistic explanations with these models". The sceptic is correct in her challenge: mechanistic commitments in terms of modeling do not imply mechanistic commitments in terms of explaining. In order to overcome this problem, I intend to show that there is a relation between engaging in mechanistic modeling and intending to explain mechanistically. I resort to an argument by Levy (2013) to make this point.

The initiative of using mechanistic modeling strategically resembles a thesis by Levy, which he calls "Strategic Mechanism" (SM). Levy's SM thesis posits that mechanistic modeling is – at least for some phenomena – the best methodological approach. The reasons why mechanistic modeling can be so convenient amount to its cognitive and epistemic contributions to research. As Levy claims, scientists – as humans – "find piecemeal, relatively concrete, visualizable representation and analysis epistemically congenial" (105). In particular, I have argued above that mechanistic modeling in the case of the BK model facilitates analysis and manipulation. And BK explicitly say that their model is mechanically convenient. In this sense, BK' efforts embody SM.

As a corollary of SM, Levy submits that mechanistic modeling involves more than just the production of mechanistic representations, namely mechanistic reasoning concerning the target of the model (cf. Levy, 2013: 107). Levy uses this corollary to explore the relations between the SM thesis and a second thesis, which he calls "explanatory mechanism" (EM).

This second thesis states that, in order to explain a phenomenon, mechanistic information must be provided. SM and EM hold close relations to each other, especially under an epistemic conception of explanation. A scientist who endorses EM, under an epistemic view of explanation, focuses on "how mechanistic information is represented in explanation" (108). If scientists endorse EM, then SM stands as a convenient companion. This is the case because SM is precisely concerned with methods of mechanistic representation and associated reasoning and investigation. To be sure, there is no necessity in combining EM with SM, but rather practical considerations.

The reverse argument follows a similar reasoning. If scientists endorse SM – like BK do – it is likely that they value the cognitive and epistemic contributions of representing and reasoning about their targets in mechanistic terms. In this sense, BK's mechanistic modeling is not a fortunate accident, but rather a reflective commitment to mechanistic reasoning concerning their target. By being committed to SM, BK are implicitly committed to achieving explanations shaped by mechanistic reasoning and investigation. It is in this sense that I suggest that BK endorse EM. In other words, BK's explanatory commitments can be safely classified as mechanistic.

Having tested the relations between explanatory commitments and accounts of explanation in the context of the BK case, I proceed now to discuss cases in which such relation is not straightforward. The core of this argument is that explanatory practices are not always accurately described at the level of accounts of explanation. I give examples that illustrate the problems of describing explanatory practices at the level of accounts of explanation. I conclude that accounts of explanation are to be used as a way to classify explanatory practices. Such classificatory endeavors often involve dismissing some of the nuances of the explanatory practice.

4.3 Explanatory Commitments beyond Accounts of Explanation

As I have argued above, accounts of explanation are philosophers' attempt to capture the coherent sets of explanatory commitments that guide scientists' explanatory practices. But note that not every coherent set of explanatory commitments that guides scientists'

explanatory practices has a single corresponding account of explanation. Or, more precisely, not all coherent sets of explanatory commitments can be straightforwardly classified as a well-established account of explanation in the philosophical literature. This is the thesis I proceed to discuss.

As a motivation, consider some views delivered by Weber *et al.* (2013). They present a table in which six "positions" can be attributed to six distinct accounts of explanation in different configurations. In this case, the accounts of explanation are identified with the name of their advocates (Table 4-1). The six analyzed positions simply are positions concerning explanatory commitments: they are guidelines that shape explanatory practices. Note that the six accounts of explanations studied by Weber *et al.* do not exhaust all the different possible configurations for these positions. In this sense, new coherent sets of commitments can be constructed, which do not necessarily align with a single well-established account of explanation in the philosophical literature.

	Hempel	Hausman	Cartwright	Humphreys	Kitcher	Salmon
Are explanations arguments?	Yes	Yes	No	No	Yes	No
Can explanations contain accidental generalisations?	Yes	No	N/A	N/A	No	N/A
Can explanations contain irrelevant premises?	Yes	No	N/A	N/A	No	N/A
Are explanations symmetrical?	Yes	No	No	No	No	No
Do explanations cite causes?	No	Yes	Yes	Yes	No	Yes
Do explanations increase the probability of the explanandum?	Yes	Yes	Yes	No	Yes	No

Table 4-1: Positions of various authors with regards to aspects of scientific explanations, from Weber et al. *(2013: 12).*

The idea that explanatory commitments can be assembled in coherent sets that go beyond the available repertoire of accounts of explanation can be further analyzed by focusing on two aspects of explanatory practices. First, explanatory practices across research programs overlap in some commitments. Based on this intersection of commitments, these practices may be classified in terms of more than one account of explanation. I refer to this as the issue of overlaps. Second, explanatory practices may integrate explanatory

commitments that are traditionally attributed to distinct accounts of explanation. I refer to this as the issue of integration. To visualize this, see Figure 4-1.

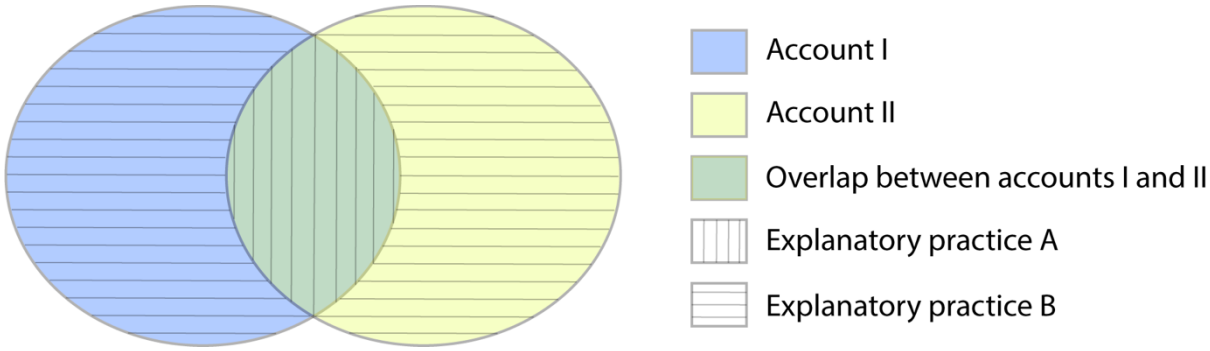


Figure 4-1: Account I and account II are distinct accounts of explanation. Explanatory practice A is guided by commitments at the intersection between accounts I and II. Explanatory practice B is guided by commitments that are integrated from accounts I and II.

I tackle first the issue of overlaps. To explore this issue, consider the various research programs engaged in explanatory practices that involve commitment to citing causes in their explanations. Philosophers of science have delivered various accounts of explanation that refer to such commitment. These accounts are referred to as "causal" accounts of explanation. For example, consider the interventionist account (Woodward, 2003), the conserved quantity account (e.g., Dowe, 1992), the kairetic account (Strevens, 2008), and the new mechanist account (e.g., Glennan, 2017), to name a few. These accounts share commitment to causation, even though they differ in other commitments and the way in which they relate to each other.

To make this more explicit, consider the differences between the conserved quantity and the new mechanist accounts. On the one hand, the conserved-quantity account – as its name suggests – relies on citing causal interactions in which exchanges of conserved physical quantities occur. On the other hand, the new mechanist account relies on the description of entities engaged in organized activities that together are responsible for a phenomenon. These accounts share commitment to the explanatory power of causal relations. In fact, one might argue that the conserved quantity account is a borderline case of the mechanistic account: the exchange of a conserved quantity can be described as an organized activity between entities. However, mechanistic explanations are commonly crafted in terms that

transcend the terminology of the conserved quantity account which is strongly grounded on physics. For example, MDC (2000) mention as examples of mechanistic activities "fitting, turning, opening, colliding, bending, and pushing." In some cases, these activities might be expressed in a way that makes explicit the conservation of some physical quantity. But that is beyond the point. Mechanistic explanations that rely on these notions are adequate according to mechanistic explanatory commitments, without mentioning the conservation of physical quantities. In this sense, the conserved quantity and the new mechanist accounts share commitment to citing causal relations, but do not share commitment on the existence of activities other than the exchange of physical quantities.

Or consider accounts of explanation that do not rely on causation, i.e., "non-causal" accounts of explanation. Non-causal accounts overlap in finding explanatory power in elements other than causal relations. Although non-causal accounts of explanation share this commitment, they differ in others. Most notably, non-causal accounts differ in their ontologies, grounding their explanatory power on different kinds of things. For example, non-causal explanation may cite structures, templates, topologies, mathematical relations, and the like.

The overlaps in explanatory commitments can be assessed in the context of Table 4-1. For example, consider commitment to citing causes in explanations. According to Weber *et al.*, Hausman, Cartwright, Humphreys and Salmon's accounts of explanation overlap in capturing an explanatory commitment to citing causal relations. In this sense, they are causal accounts. Although these accounts share commitment to citing causal relations, they differ in other commitments. For example, Humphreys does not accept that increasing the probability of an explanandum should be part of what counts as an explanation. He is not committed to this consideration. In contrast, the other three causal accounts are so committed to this consideration.

Now, consider the issue of overlaps in the context of my case studies. As I argued above, two explanatory commitments in the BK case are discreteness and modularity. I suggest that these explanatory commitments are also part of the OFC case. Indeed, OFC engage in the construction of a model which is – like the BK model – discrete and modular,

namely a cellular automaton. This is an overlap in explanatory commitments. To put this in the context of Table 4-1, if OFC and BK were asked "are your explanations crafted in terms of discrete entities" and "do your explanations resort to modularity," they would answer "yes" to both of these questions.

However, as opposed to the BK case, I do not suggest that a mechanistic account of explanation is the best way to classify OFC's explanatory commitments. This is because the kind of explanation that OFC aim for is crafted at the level of mathematical abstractions. That is, although BK and OFC overlap on the commitment to discreteness and modularity, they differ in other commitments. This difference makes BK's explanations best classified as mechanistic and OFC's explanations best classified as mathematical.

The differences between BK and OFC's explanatory commitments can be clearly attested in the implementation of their models. The BK model, in its laboratory implementation, is a concrete mechanism, i.e., a set of entities (blocks and springs) engaged in organized activities. In its mathematical implementation, the BK model is a physico-mathematical description of the spring-block system. It explicitly resorts to physical notions such as "mass", "elastic constant", and "velocity", to name a few. The resulting description is written as a system of mathematical equations, but these equations represent the physics of the spring-block system in the lights of Newtonian physical theory. In contrast, the OFC model does not resort to physical notions in its descriptions. The OFC model is a set of variables (the cells in the cellular automaton) and a set of instructions that establish the evolution of these variables. The instructions amount to simple arithmetic operations among variables, and the evolution of variables consists in changes in the numerical value attached to them.

There is a sense in which the instructions in the OFC model can be thought of as representing physical phenomena in the BK model. In fact, this is how OFC interpret them. For example, the existence of a static friction threshold in the BK case is analogous to the thresholds defined in the cellular automaton in the OFC case. Or, the transmission of elastic tension to neighboring cells in the BK case is analogous to the subtractions and additions in contiguous cells in the cellular automaton in the OFC case. Beyond these analogies, my point is that, in the OFC case, the instructions are not coded as mathematical representations of

physical theory. In other words, OFC's cellular automaton is not solving physics. Rather, OFC's cellular automaton is a mathematical calculus that resembles – in a highly abstract degree – the physics of a spring-block system.

Now, consider the second aspect in which explanatory commitments may go beyond accounts of explanation, namely the issue of integration of explanatory commitments. Scientists' explanatory practices might deploy various explanatory commitments that are not always univocally attributed to a single, well-established account of explanation. The fact that scientists are able to put together these explanatory commitments as part of their research means that these explanatory commitments cohere in the context of the program. This is the case even though the explanatory commitments may traditionally be attributed to distinct accounts of explanation. In this sense, I suggest that accounts of explanation amount only to stable and robust sets of explanatory commitments that philosophers use to describe scientists' explanatory practices. But scientists are not restricted to accepting and assembling sets of explanatory commitments that coincide with philosophers' accounts of explanation.

Naturally, philosophers of science could come up with new accounts of explanation that attempt to reflect the integration of explanatory commitments. A good example of this is Strevens's kairetic account. Strevens (2004) argues that his kairetic account unifies the causal and unification accounts of explanation. I take Strevens's efforts to be an attempt to integrate explanatory commitments that are idiosyncratic to two distinct accounts of explanation. As he expresses in the abstract of his paper, he takes "the central technical apparatus of the unification account to solve a central problem faced by the causal account, namely, the problem of determining which parts of a causal network are explanatorily relevant to the occurrence of an explanandum. The result is a causal account of explanation that has many of the advantages of the unification account" (154).

In the context of my case studies, I consider the SOC program to be a good example of integration of commitments typically attributed to different accounts of explanation. I suggest that the explanatory practices in the context of the SOC program cannot be described in terms of a single account of explanation. In the following, I explore three explanatory

commitments that guide explanatory practices of the BTW group. I observe that these commitments are traditionally attributed to distinct accounts of explanation.

A first candidate, inferred from the reading of BTW's seminal paper, is causal explanatory commitments. Indeed, BTW introduce SOC as a mechanism responsible for spatial and temporal power-law correlations in complex systems: "We suggest that this self-organized criticality is the common underlying mechanism behind the phenomena described above" (Bak *et al.* 1987: 381). The explanandum is spatial and temporal power-law correlations. The explanans is the SOC mechanism. The relation between explanandum and explanans is one of causal dependence. As a caveat, the notion of mechanism employed by BTW departs from the new mechanist sense. Rather, the term 'mechanism' refers to a set of causal factors. This set of causal factors is discussed in Section 2.3.2 as the SOC genotype (Watkins *et al.* 2016).

A second candidate is unificationist explanatory commitments. The SOC program has often been characterized as a unificationist endeavor. This is an explicit motivation for several scientists working in the program, but specially so for BTW. Arguably, one of the most striking expressions of such unifying attitude is shown in Bak's 1996 book *How Nature Works*. According to this unificatory attitude, the various natural phenomena manifesting spatial and temporal power-law correlations are explained by providing a single theory that accounts for their occurrence. Arguably, this unificationist attitude arises from the observation that SOC seems ubiquitous. As I argued above, the ubiquity or generality of SOC in nature is a controversial feature to assert. Despite the problems with the "ubiquity" claim, it is undeniable that such stance did impact the SOC program and helped organize a great part of the efforts. As an example, consider the following quote by BTW (1988):

> *"This paper concerns the behavior of spatially extended dynamical systems —that is, systems with both temporal and spatial degrees of freedom. Such systems are common in physics, biology, and even social sciences such as economics. Despite their abundance, there is little understanding of the spatiotemporal evolution of these complex systems. Seemingly disconnected from this problem are two widely occurring phenomena whose very generality require some unifying underlying explanation. The first is a temporal effect*

known as 1/f noise or flicker noise; the second concerns the evolution of a spatial structure with scale-invariant, self-similar (fractal) properties" (364; my emphasis).

To be sure, this unificatory attitude does not exclude causal explanatory commitments. In fact, unificatory explanatory commitments are compatible and often explicitly stated in tandem with causal explanatory commitments. For example, the SOC mechanism is introduced as "the *common* underlying *mechanism* behind [power-law temporal and spatial correlations]" (BTW, 1987: 381; my emphasis). It is in this sense that I suggest that explanatory commitments can be integrated in coherent sets, even though they are not straightforwardly attributed to a single account of explanation.

The third candidate is mathematical explanatory commitments. These commitments are grounded in one of the main methods in the SOC program: computer simulation with cellular automata. As discussed above, cellular automata are abstract mathematical constructs (Section 2.3.3). In this sense, the features of a cellular automaton, its outcomes, and the relations between features and outcomes are a subject matter of mathematical research. If approached in this way, cellular automata can be studied in isolation from their representational capacities. In other words, cellular automata may themselves be the object of research. This is, indeed, an important field of research. As Toffoli and Margolus (1990) assert, cellular automata are "abstract dynamical systems that play a role in discrete mathematics comparable to that played by partial differential equations in the mathematics of the continuum" (229). In this sense, explaining the outcomes of a cellular automaton simulation is a problem in mathematics.

These explanatory commitments are integrated in the SOC program, even though they are traditionally assigned to distinct accounts of explanation. Because of this, using a single account of explanation as a descriptive tool potentially misses some nuances of the explanatory practice. Furthermore, the way in which these explanatory commitments are integrated in explanatory practice is susceptible to change. In particular, the status of these commitments as core or peripheral may evolve and, with them, the resulting explanatory practices. For example, it is interesting to note that unificationist commitments – while being central in the original SOC program – have become more peripheral with the evolution of the

program. As a result, current explanatory practices – for the most part – are not so much focused on stating a common mechanism for SOC phenomena. Rather, the current scope of explanations in SOC is better constrained to either abstract mathematical claims or particular physical phenomena.

The issues of overlaps and integration of explanatory commitments bring me to my reservations about Weber *et al.*'s project of accounts of explanation as a toolbox (2013). For them, the various accounts of explanations serve as tools for the description and evaluation of explanatory practices in specific contexts. I think this project is a step in the right direction, in the sense of embracing the plurality of accounts of explanation and putting them to work in context. Nonetheless, I consider that their emphasis on accounts of explanations prompts them to provide inaccurate descriptions and evaluations in cases where overlap and integration of explanatory commitments are relevant.

Consider the descriptive side of their project. Often, the description of explanatory practices in terms of accounts of explanation is – I suggest – too coarse-grained. In practice, describing explanatory practices in terms of accounts of explanation amounts to a classification. As any classificatory enterprise, nuances are missed in the process. I suggest that a finer-grained description, at the level of explanatory commitments, provides a more accurate description of the explanatory practices. This is not to say that classification is not a valuable task. In fact, I do engage in classificatory efforts, e.g., I classify BK's explanatory commitments as mechanistic. My point is that classification in terms of an account of explanation should be disentangled from a fine-grained description. This is why I first describe BK's explanatory commitments in terms of discreteness and modularity and only then attribute them to the new mechanist account of explanation.

With regards to the evaluative side of their project, I also consider it inadequate. Weber *et al.* emphasize the practical nature of evaluating explanatory practices. They argue that the evaluation of explanations should take into account scientists' epistemic interests (35). An explanation is appropriate only relative to the achievement of such epistemic interests. Subsequently, Weber *et al.* argue that they are open to using all available accounts

of explanation – i.e., the toolbox – to evaluate explanatory practices. In this sense, they describe their approach as methodologically neutral and implicitly pluralistic.

The problem – I suggest – appears in their next step. Weber *et al.* argue that, once recognized the epistemic interests – expressed in the form of scientific questions – one or another account of explanation is more appropriate. I think this approach is not a good solution precisely because of the issues of overlaps and integration. Scientists' epistemic interests may be at the intersection of various accounts of explanation. Or scientists' epistemic interests may integrate elements of various accounts of explanation. Because of this, the level of evaluation should not be that of accounts of explanation.

In the next chapter, I continue discussing explanatory commitments but in a different setting. The next step of my argument is to argue that explanatory commitments play a central role in the interpretation of objects that are intended to be used as models. This is particularly the case when the model is intended to be used for model explanations. Depending on their accepted explanatory commitments, scientists can deliver various interpretations of such objects.

4.4 Précis

The main theoretical theses of this chapter are:

- Thesis 1: Explanation may be conceived ontically, epistemically, or as a combination of both to some degree.
- Thesis 2: Accounts of explanation (in the scientific sense) are coherent bundles of explanatory commitments which guide explanatory practice.
- Thesis 3: Accounts of explanation (in the philosophical sense) are philosophers' attempt to articulate accounts of explanation (in the scientific sense).
- Thesis 4 (the issue of overlaps): Explanatory commitments that guide explanatory practices may be captured at the intersection of distinct accounts of explanation.
- Thesis 5 (the issue of integration): Explanatory commitments in explanatory practices may integrate commitments that are idiosyncratic to distinct accounts of explanation.

- Thesis 6: Explanatory practices are more accurately described at the level of explanatory commitments. These commitments can be classified in terms of accounts of explanation, but such classification misses some of the nuances of a fine-grained description.

The main lessons learnt from my case studies are:

- Explanation is a central aim of the SOC program. In particular, it is a central aim in the OFC model. It is also an important aim in the BK case, even though they are less explicit about it.
- The SOC program is committed to both ontic and epistemic views on explanation. However, the ontic motivations have been thoroughly questioned.
- BK's explanatory commitments can be adequately described as mechanistic (e.g., discreteness and modularity in the spring-block model).
- OFC share commitment to discreteness and modularity. However, in the OFC case, these commitments are better classified as mathematical commitments.
- The SOC program integrates explanatory commitments that are traditionally attributed to distinct accounts of explanation, namely unificationist, causal, and mathematical explanatory commitments.

***

5 Interpretation in Models

In this chapter, I investigate the notion of interpretation. Broadly speaking, interpretation amounts to bestowing meaning upon stimuli. Human agents commonly bestow meaning to a wide variety of stimuli. These go from linguistic stimuli – such as utterances and written words or sentences – to non-linguistic stimuli – such as concrete objects, images, and body movements. In addition, humans bestow meaning to stimuli in diverse contexts which affect the resulting interpretation. These contexts may go from artistic endeavors to political discourse.

These two dimensions of variation – i.e., the object and context of interpretation – make the topic of interpretation overwhelmingly broad. For the purposes of this dissertation, I set restrictions to my treatment of the topic of interpretation in terms of these two dimensions. First, I focus on studying the topic of interpretation in a particular context, namely that of scientific inquiry. I engage with case studies of scientific research, namely the BK and OFC cases. I also engage with the literature on interpretation in the domain of philosophy of science. Fortunately for my project, a great deal has been written on interpretation within this domain. Thus, I do not attempt to engage with the vast literature on hermeneutics, semiotics, and the like. This does not mean that I do not engage with ideas derived from these disciplines. It is clear that they have had a significant impact in the philosophy of science. My point is that I relate to them only to the extent in which they have permeated through the philosophy of science into those discussion in which I am interested. Second, I focus on one kind of object of interpretation, namely objects that are used as scientific models. The reason for this choice is the central role that scientific models play in my argument. My main concern is to explore how models are used to craft explanations. More explicitly, I am interested in explanations that make reference to the content of a scientific model. These explanations are often referred to as "model explanations."

The role of interpretation in my overall argument is the following. In order to use an object as a scientific model, it must be bestowed with meaning. In this sense, a scientific

model is an interpreted object. When used for explanatory purposes, the interpretation embodied in models is guided by explanatory commitments. In this sense, model explanations rely on the interpretations that are embodied in a given model. As Faye (2014) points out: "[B]oth interpretation and explanation are parts of the same scientific practice" (64). Scientific models may embody different interpretations, depending upon the different explanatory commitments held by the interpreter scientist (or group of scientists). These various interpretations allow for the crafting of distinct model explanations with different content.

The structure of this chapter is as follows. First, I discuss two preliminary issues: I clarify the senses in which I use the notions of "interpretation" and "scientific models." Second, I present two modes of interpretation concerning models, namely non-representational and representational interpretation. I illustrate and explore these modes of interpretation in the context of my case studies. Third, I discuss the relation between commitments and interpretation. In particular, I focus on the relation between explanatory commitments and non-representational interpretation.

5.1 Preliminaries on "Interpretation" and "Scientific Models"

Interpretation can be conceived as the bestowing of meaning to a stimulus, with 'stimulus' broadly construed. There are different approaches to bestowing meaning to stimuli. One could think of stimuli as having meaning in themselves, a meaning that is objective and can be discovered. Or one could think of stimuli as having no intrinsic meaning. In this case, meaning must be put there by an interpreter. Or one could think of something in between: stimuli have a meaning that is not intrinsic to themselves, but neither fully relative to the interpreter. Rather, the meaning is intersubjectively decided, i.e., based on a community of interpreters. Something along the lines of the latter approach has been referred to as "pragmatic" approach to interpretation (cf. Dascal, 2003). This is the approach I assume in this chapter and for the most part of this dissertation.

The pragmatic approach to interpretation posits interpretation as an activity in which agents engage. However, the term 'interpretation' is also often used to refer to the outcome

of such activity, i.e., the bestowed meaning. As an illustration, I consider dictionary entries of 'interpretation'. Lexico defines 'interpretation' as: i) "the action of explaining the meaning of something"; and ii) "an explanation or way of explaining." It is noteworthy that the first sense posits 'interpretation' as an action, while the second sense posits 'interpretation' as the outcome of such action. In this dissertation, I embrace this means-end dichotomy. That is, I take interpretation to be both an activity and its outcome, unless specified in a particular context.

The main context in which I intend to apply the notion of interpretation is the context of scientific research with models. Because of this, I spend some time clarifying the notion of scientific model in the context of this dissertation. This is a difficult task. The topic of scientific models has become one of the main endeavors in the philosophy of science in the last decades. This should come as no surprise: scientific models play central and heterogenous roles in scientific research. As a result, the philosophical literature on scientific models has become not only vast, but also diverse in their claims. It is not my intention to review this literature.[39] Instead, I intend to present a minimal characterization of "scientific models" which addresses some of the platitudes found in the literature.

I submit that a scientific model can be minimally characterized as an interpreted vehicle which is used for scientific purposes.[40] I proceed to unpack this characterization in its three main components: i) vehicle; ii) scientific purposes; and iii) interpretation. First, the term 'vehicle' refers to an object that is used as a model. I adopt the term from papers such as Frigg and Nguyen (2016) and Contessa (2007). Other terms with a similar connotation are used elsewhere. For example, Suarez (2015) uses the term "model source" and Frigg and Nguyen (2018) use the term "base" (in the case of a material vehicle). In general, a vehicle is merely an object. In fact, any kind of object has the potential to be used as a model. In particular, vehicles can be concrete objects (such as those in scale models or model organisms) or abstract ones (such as mathematical entities or fictional settings).[41] As Suarez

[39] For reviews of this sort, see Frigg and Hartmann: 2020; or Magnani and Bertolotti: 2017.
[40] This characterization resembles that of Frigg and Nguyen (2016) who claim that a model M is a duple $\langle X, I \rangle$, where X is a vehicle and I is an interpretation.
[41] Note that I use the term 'abstract' in opposition to 'concrete'. Later in this dissertation, I use the term 'abstract' in opposition to 'complete', in the context of discussions on idealization.

(2015) claims: "[Vehicles] may be concrete or abstract, physical or mathematical, real or imaginary" (41). In this sense, the debates on the ontology of models can be cashed out as debates on the ontology of vehicles that are used as models. In discussing the ontology of models, Frigg and Hartmann (2020) mention that models may be physical objects, fictional objects, set-theoretic structures, descriptions, equations, and gerrymandered objects (i.e., combinations of the previous categories).

As objects, vehicles have an associated set of properties. These properties are said to be instantiated by the vehicle (Frigg & Nguyen, 2016: 231). Instantiated properties are constrained by the kind of object that a vehicle is. For example, physical objects may instantiate a color and mass, while set-theoretic structures do not instantiate these properties. For the purposes of this dissertation, I assume that scientists are able to individuate a vehicle and study its instantiated features. This is a level of analysis in which my argument bottoms out, by positing the study of vehicles and their properties as unproblematic.

As an illustration, consider my case studies. The BK model has – in fact – two distinct vehicles, one for each implementation. In the laboratory implementation, the vehicle is a concrete physical object, namely the spring-block system. In the mathematical implementation, the BK vehicle is an abstract object, namely a set of ordinary differential equations. The OFC model has as vehicle an abstract (mathematical) object, namely a cellular automaton.

It is not uncontroversial to say that a system of ordinary differential equations or a cellular automaton can be taken as the vehicles of the BK and OFC models, respectively. Criticism has been directed towards approaches that admit abstract objects as vehicles of models. For example, Knuuttila and Voutilainen (2003) argue for the materiality of models. Roughly, their point is that models are materialized by the concrete practices of scientists who employ them. Knuuttila and Voutilainen have a point. Even in the case of models traditionally conceived as abstract, such as mathematical models or fictional models, scientists relate to them by means of concrete practices. This approach may well be applied to my case studies. In this view, the vehicles are not the ordinary differential equations and

the cellular automaton. Rather, the vehicles are the computer programs in which the mathematical apparatus is implemented, or the set of practices surrounding the mathematical research. For the moment being, this debate does not have a major impact in my overall argument. Thus, I remain open to both readings.

Moving on to the second component of my characterization, scientific models are used for *scientific purposes*. As Gelfert (2017) proposes, scientific models are "functional entities." But they are functional entities in the domains of science. This second aspect of my characterization is intended to distinguish scientific models from those interpreted vehicles that are used for non-scientific purposes. Consider for example the interpretation of movies, texts, facial expressions, theater plays, or the stars (in the astrological sense). All of these are vehicles that can be interpreted, but the guiding motivations need not be scientific (e.g., religious, political, and aesthetical purposes). In this sense, they are not scientific models.

Scientific purposes are diverse. Some of the typical ones that prompt the usage of models are (more or less accurate) descriptions of phenomena, (more or less accurate) predictions, generation of (more or less sound) inferences concerning a phenomenon, and (various kinds of) explanations, just to name a few common ones. In this dissertation, I mostly focus on the latter purpose, namely in the form of model explanations. Later in this dissertation, I relate the attainment of model explanations to the achievement of understanding (Chapter 7).

Using my case studies as an illustration, it is safe to say that the BK model and the OFC model are used for scientific purposes. In the previous chapter, I argue that both cases aim for explanations of the occurrence of earthquakes. For example, OFC claim that "[t]he model gives a good prediction of the Gutenberg-Richter law and an explanation to the variances in the observed b values" (1244). BK also declare that the BK model was constructed to "explore the role of friction along a fault as a factor in the earthquake mechanism" (341). This exploration responds to scientific research in seismology, and it is published in the Bulletin of the Seismological Society of America. In the OFC case, they frame their research within a scientific research program, namely the SOC program, and their work is published in the Physical Review Letters.

Third, vehicles must undergo a process of *interpretation* to be used as scientific models.[42] That is, they must be bestowed with meaning by the scientists using them as models. The general idea is that vehicles are used as tools to attain scientific purposes. But the functionality of a vehicle as a tool – although constrained by its nature – is not intrinsic to it. A vehicle has the potential to be used in different ways to advance the attainment of various scientific purposes depending on the interpretation that is given to it. To be sure, the fact that a vehicle has no intrinsic function and may be subjected to various interpretations does not mean that it can be submitted to just any interpretation. Vehicles instantiate properties which constrain the legitimate interpretations and uses to which they can be subjected. In other words, they have constraints and affordances (cf. Knuuttila & Voutilainen, 2003: 1487). In the next section, I proceed to analyze the interpretation of vehicles in more detail. In particular, I distinguish two modes of interpretation, namely non-representational and representational.

5.2 Non-Representational and Representational Interpretation

To motivate the discussion for this section, I reflect upon the role of a picture. A picture is an object conventionally used for representational purposes. That is, the typical role of a picture is to depict something. Accordingly, a common way to attribute meaning to a picture is by deciding what it depicts and how it does so. This is an instance of what I refer to as a "representational interpretation." In this case, the interpretation of the picture is guided by the intended representational role of the picture. Before such interpretation, the picture has no meaning whatsoever. Consider this quote by van Fraassen (2008): "If we were to ask 'What is in a picture?' while taking the picture simply to be the physical object and with no relation to anything that can bestow meaning, the answer would have to be 'Nothing!'" (25).

[42] In a famous quote by von Neumann (1955), he argues that "[t]he sciences do not try to explain, *hardly even try to interpret*, they mainly make model" (492; my emphasis). In contrast, I submit that the making of models requires interpretation. Von Neumann himself seems to admit this later in the same passage: "By a model is meant a mathematical construct which, with the addition of certain verbal *interpretations* describes observed phenomena" (ibid; my emphasis).

Although representational interpretations are commonplace in bestowing meaning to pictures, I submit that this is not the only possible approach to attribute meaning to them. More explicitly, pictures can be used in ways other than as representational devices. For example, consider that a picture – that is, the physical object – can be simply interpreted as a piece of paper and used to do an origami with it. In this case, the picture is not employed as a representation, and its interpretation as a foldable piece of paper has no bearing on its representational force. This is what I call a "non-representational interpretation," i.e., bestowed meaning that has no bearing on the vehicle's representational status.

Leaving pictures aside, I now consider the same issues in the context of scientific models. Scientific models often work as pictures: they are typically used as a representation of a target phenomenon. This use requires a representational interpretation of a vehicle. That is, meaning is bestowed on a vehicle in terms of its target and how it relates to it. In other words, the meaning of a vehicle in representational mode is its *denotatum* (i.e., what is being denoted). In addition, scientific models can be used as objects of study in themselves with no representational function. This use involves a non-representational interpretation of a vehicle. Roughly speaking, non-representational interpretation amounts to conceptualizing a vehicle and schematizing it, i.e., distinguishing its parts, properties, and relations.

Given that a scientific model is an interpreted vehicle, different interpretations of a same vehicle generate different models. This statement holds for both non-representational and representational interpretations. Consider representational reinterpretation. Vehicles can be reinterpreted in terms of targets different from the one that originally motivated the vehicle's selection, design or construction. Contessa (2007) discusses this issue in terms of the "retargeting" of models (62). But reinterpretations can also be non-representational. That is, a same vehicle can be conceptualized or schematized differently. As an illustration, the OFC vehicle – i.e., a cellular automaton – can be non-representationally interpreted in various ways. For example, the OFC vehicle may be interpreted as an abstract mathematical object, as the corresponding algorithm that describes its dynamics, the particular code of a computer program, or the physical processes occurring in the computer simulation.

The distinction between representational and non-representational interpretation has been discussed elsewhere in the literature with different terminology. I briefly comment on four such cases due to van Fraassen (1994), Faye (2014), Giere (1988) and Frigg and Nguyen (2016). For example, consider van Fraassen (1994). He claims that "in science too [i.e., together with the fine arts], we find interpretation at two different levels: the theory represents the phenomena *as* thus or so, *and* that representation itself is subject to more than one tenable but significantly different interpretation. The texts of science too are open texts" (177). The first level that van Fraassen addresses is that of representational interpretation. In this case, he focuses on theories as the vehicle of representation (but, as I have argued, several kinds of vehicles can be used for representational purposes in science). The second level is that of non-representational interpretation: the level at which the vehicle of a representation itself can be subjected to different interpretations, regardless of its *denotata*.

Faye (2014) introduces a distinction between "determinative" and "investigative" interpretation, which resembles that of representational and non-representational interpretation (61-2). The former is concerned with the explanation of meaning, while the latter is concerned with the construction of meaning. In the determinative case, the interpretive question is what an object stands for or what it refers to. As Faye (2014) says: "A determinative interpretation proposes a deliberately formulated hypothesis concerning what it is that a representation really represents" (62). In the investigative case, the interpretative question is how an object can be made a subject of representation in the first place. In other words, investigative interpretations propose a classification or conceptual representation (i.e., conceptualization) for the object of interpretation.

Giere (1988) also discusses some notions that resemble the distinction between representational and non-representational interpretation, namely the notions of "interpretation" and "identification" (75). Giere discusses these notions in the context of modeling practices with mathematical models. In this context, interpretation is the "linking of the mathematical symbols with general terms, or concepts […]" (ibid). In contrast, identification is the "linking of a mathematical symbol with some feature of a specific object" (ibid). I suggest that Giere's notion of interpretation is a form of non-representational

interpretation, in which the mathematical symbols are conceptualized. In contrast, Giere's notion of identification is a form of representational interpretation in which mathematical symbols stand for specific features of an object.

Finally, my distinction between representational and non-representational interpretation parallels that of Frigg and Nguyen (2016) between vehicles being "representations-of" and vehicles being "Z-representations" (see also "representation-as Z" in Goodman, 1976). They claim: "Being a Z-representation is a one-place predicate that categorizes representations according to their subject matter [i.e., Z]; being a representation-of is a binary relation that holds between a symbol and that which it denotes. The two can, but need not, coincide [i.e., a representation-of Z may also be a Z-representation]" (Frigg & Nguyen, 2016: 227). The parallels are the following: For a vehicle to be a representation-of, it must be representationally interpreted. For a vehicle to be a Z-representation, it must be non-representationally interpreted.

Representational and non-representational interpretation can be identified in my case studies. As an illustration, I consider the BK case. In some passages and sections of their paper, BK's interpretation of the vehicles is non-representational. In the passages in which they discuss the laboratory implementation, BK focus on describing and analyzing the dynamics of the BK vehicle *as* spring-block system. In these passages, there is no reference to the intended target, namely earthquakes in geological faults. For example, in section II of their paper, BK conceptualize the vehicle as a concrete system and identify its component parts as "masses", "springs", "rough surface", and "motor". And, in those passages in which the mathematical implementation is discussed, BK engage in a mathematical study of the system of differential equations. In these passages, the equations are studied as mathematical entities, not as representations of the dynamics of geological faults. In contrast, there are passages in which the representational component of interpretation is clearly expressed. This is salient in sections where simulation results and empirical evidence of the statistical behavior of naturally occurring seismic earthquakes are compared.

The representational component of interpretation has been widely discussed in the literature, mostly in the form of denotation and mappings. In contrast, the non-

representational component of interpretation has received comparatively little attention. In fact, several philosophers seem to reduce the value of models to their representational capacities or, at the very least, pose representation as a crucial property of models. Knuuttila (2005) reflects on this issue:

> *"Interestingly, the critics of the semantic approach nevertheless share with them the same presupposition that* the main task of models is to represent the world*. In arguing against the semantic view of models,* some of them have explicitly made representation the crucial property of models *(see e.g. Frigg 2003[b], 33; Hughes 1997, and Suárez 1999), which is presumably something that most philosophers generally agree in" (43; my emphasis).*

Such emphasis on the representational capacities of models is misleading in two senses. First, models do not need to be used for representational purposes. They can be used non-representationally, as objects of study in themselves (cf. Knuuttila, 2005). This means that non-representational interpretation does not require representational interpretation. Second, even if used representationally, vehicles must undergo a non-representational interpretation first. This amounts to a previous conceptualization or schematization of the vehicle in order to be used as representational model. In other words, representational interpretation necessitates non-representational interpretation.

This asymmetric relation between representational and non-representational interpretation has been noted and discussed in the literature. For example, Suarez (2015) – in discussing Hughes' DDI account – comments on two distinct activities involved in interpretation. First, interpretation "requires the ascribing of some structure to the source [i.e., vehicle] and target objects, by judiciously partitioning them into an appropriate set of parts and their properties" (44). This activity described by Suarez amounts to what I call "non-representational interpretation". Second, interpretation "also calls for a mapping of the elements of the source [i.e., vehicle] structure onto some corresponding parts and properties of the target, again *under some suitable partition*" (ibid; my emphasis). This amounts to what I refer to as representational interpretation. As suggested by Suarez, a representational interpretation (in this case expressed as "mapping") requires a non-representational

interpretation (in this case expressed as “partition”). Indeed, if the vehicle is not partitioned into parts, properties, and relations, then there is no *relata* to map to the target.

The asymmetry between representational and non-representational interpretation can be tested in the context of the BK case. As I argued above, there are passages in their paper in which BK study the vehicle as spring-block system, without discussing its representational employment. This is an instance of non-representational interpretation without representational interpretation. In contrast, representational interpretations seem to rely on a pre-existing non-representational interpretation. For example, in section VII of their paper, BK interpret the spring-block system representationally as standing for segments of seismic faults with different viscosity (Figure 2-9). However, in order to provide this representational interpretation, BK must schematize the spring-block system. This consists in distinguishing sets of blocks and assigning different laws of friction to them. The distinction of sets of blocks and assignment of specific properties to them amount to a non-representational interpretation of the spring-block system.

In the next sections, I proceed to analyze in more details the component tasks of non-representational and representational interpretation. I submit that non-representational interpretation consists of three component tasks: i) conceptualization; ii) schematization; and iii) exemplification. In contrast, representational interpretation consists of two component tasks, namely i) denotative function; and ii) mapping. I proceed first with the analysis of non-representational interpretation.

5.3 Non-Representational Interpretation as Conceptualization, Schematization, and Exemplification

The most basic form of non-representational interpretation of a vehicle is “conceptualization.” This is endorsed by Frigg (2003b), who reflects upon the foundational role of conceptualization in modeling: “Constructing a model is essentially a process of conceptualisation. One begins with some visual and experimental evidence about the behavior of the system (in some cases also some background theory) and then interprets, conceptualises, and categorises this evidence in a certain way” (627).

Roughly speaking, conceptualization means representation by means of a concept (cf. "conceptual representation" in Faye, 2014: 61-2). Even though conceptualization involves representation, it should not be conceived as a task of representational interpretation. In representational interpretation, a vehicle becomes a representation *of* a target. In conceptualization, the vehicle becomes represented *by* a concept. In more technical terms, in a representational interpretation, a vehicle becomes a *representans* for a *representatum* target phenomenon. In contrast, in non-representational interpretation (conceptualization in particular), the vehicle becomes a *representatum* of a *representans* concept.

A wide variety of concepts may be used to conceptualize a vehicle. As a toy example, consider an object, say a person. This object may be conceptualized not only as a person (obviously), but also as a human being, a mother, a collection of organs, a single unit, a potential customer, a mass of 80 kg, just to name a few examples. Below, I engage in a deeper discussion about the considerations that shape the choice of concepts in the conceptualization of vehicles. Still, the gist of it – as the reader might expect – is that scientific commitments have a major say in the way scientists conceptualize vehicles intended to be used as models.

In addition to conceptualization, non-representational interpretation also involves a fine-grained mapping between vehicle and concept, to which I refer as "schematization." To explicate this fine-grained mapping, I resort to some views due to Frigg and Nguyen (2016) and adapt them to my terminology. The main idea is that vehicles can be partitioned in a way that the component parts of the partition can be assigned to the component parts of the concept. This presumes that concepts can be construed as having an inner structure, based on a collection of subconcepts (see the "classical theory of concepts" in Margolis & Laurence, 2019). In other words, schematization provides the resulting model with a structure based on the component parts of the concept used in its conceptualization.

Frigg and Nguyen discuss the mapping between vehicle and concept as the assignment of properties of the vehicle to component properties of a concept, with 'property' broadly construed. For example, 'properties' might include relations, structures, features, parts, and

the like. Their idea is the following: Consider a vehicle V and a set of instantiated properties $V = \{V_1 \ldots V_n\}$ (V-properties). Consider a concept Z and a set of its properties $Z = \{Z_1 \ldots Z_n\}$ (Z-properties). A bijective mapping (call it I) can be established from V-properties to Z-properties. The partitioning of V into V-properties and its mapping to Z-properties is what I call a "schematization" of V. After conceptualization Z and schematization I, the vehicle V becomes a model M which is said to I-instantiate Z-properties.

In practice, from all the Z-properties I-instantiated by M, only a subset is the focus of scientists. That is, scientists are selective in their approach to models. This brings me to the third component task of non-representational interpretation: "exemplification." I adapt some views by Elgin (2017: Chapter 9) to my views. The main idea of exemplification is that a model M I-instantiates several Z-properties, but it I-exemplifies only those that are highlighted. To highlight an instantiated feature roughly means to point at it and exclude distractors. In other words, highlighting is emphasizing. Goodman (1976) uses similar notions such as "exhibiting", "typifying", "to show forth" (86). Exemplification requires a previous conceptualization and schematization. Also, given a fixed conceptualization and schematization, different exemplifications may be established.

As an illustration, consider the BK case. I attempt to reconstruct BK's non-representational interpretation of one of their vehicles, namely the concrete object employed in the laboratory implementation. First, the vehicle is conceptualized as a 'system'. This is expressed in some passages in which BK describe the laboratory simulation:

> *"Let the leading mass be connected through a spring to a source of force which pulls the entire* system *longitudinally. Let the entire* system *rest on a stationary rough horizontal plane surface. Let the last mass be free. It is clear that in this* system *there is permanent deformation of the* system *of masses and springs relative to the surface on which the masses rest, just as there was in the first example" (344; my emphasis).*

Second, the vehicle is schematized in accordance with the concept of 'system'. As a concept, 'system' has an inner definitional structure which includes subconcepts like

'component parts' and 'interactions'. For example, Lexico defines 'system' as a "a set of things working together as parts of a mechanism or an interconnecting network" (note the relations between 'system' and 'mechanism'). The vehicle is thus schematized by identifying component parts and interactions. Roughly, the component parts are the blocks and springs, and the interactions are the elastic forces between blocks. Alternatively, one could argue that the vehicle is conceptualized as something more specific than 'system', namely a 'spring-block-system'. Accordingly, its schematization consists in identifying springs and blocks and interactions among them.[43]

Third, some features of this system are exemplified in the BK model. For example, BK highlight the frictional forces between blocks and surface, the elastic forces among blocks, the velocity of pulling, the masses of blocks, to name some of the most salient features. Note that there are features that are instantiated by the spring-block system but are not exemplified. For example: the friction between blocks and air, the size of the blocks, the orientation of the pulling, and the color of the blocks, to name just a few.

I have presented these three tasks of non-representational interpretation – conceptualization, schematization, and exemplification – as if they followed a lineal path. That is, conceptualization enables schematization, and both together enable exemplification. Although this might be a pedagogical way to present my account, it must be qualified. Conceptually speaking, these three tasks can be individuated, and each can be taken as a requisite for the next one. However, in practice, these tasks cannot be sharply distinguished. In fact, scientists might prefer one or other conceptualization because one or the other is more convenient for an intended schematization and exemplification. Instead of a simple causal path between these tasks – from conceptualization to exemplification – I posit that they stand in reflective equilibrium. This recalls the reflective equilibrium among commitments in a program (see Chapter 3). This is no accident: as I argue below, interpretations are guided by commitments. In this sense, if commitments within a program

[43] Together, the tasks of conceptualization and schematization resemble Cartwright's notion of "prepared description" (1983: 133-4).

evolve in accordance to reflective equilibrium, so do the outcomes of using those commitments, outcomes such as interpretations.

A final qualification must be stated. I have discussed non-representational interpretation of vehicles as if the latter were already existing objects, ready to be interpreted. This is an expedient simplification I have assumed in order to develop my account of non-representational interpretation. In actual scientific practice, vehicles may be selected among existing objects. But they are often put together or constructed by scientists. Typically, in the case of vehicle building, non-representational (and representational) interpretations of the vehicle are implicitly operating during its construction stage. That is, scientists craft models' vehicles with specific conceptualizations, schematizations, and exemplifications in mind. In other words, scientists have preconceived applications and motivations for the vehicle to be used as a model. However, once they are constructed, vehicles become autonomous objects. That is, they are not attached to specific interpretations, either representational or non-representational. This allows for re-interpretations of vehicles, which might be different from the original interpretations in both the non-representational and representational components. This is the context in which my argument focuses: the context in which an already existing vehicle can be re-interpreted to attain distinct epistemic achievements. This point will become clearer below as I discuss the role of commitments in interpretations.

5.4 Representational Interpretation as Establishing a Denotative Function and a Mapping between Vehicle and Target

The topic of representation – and, in particular, of scientific representation – has been intensely debated in the philosophy of science literature. I avoid reviewing these debates to keep my argument focused. Thus, my strategy is simply to state my views on scientific representation, but I do so by endorsing well-established positions in the literature. This way, I locate my views in the space of the debates. Beyond merely stating my views, I also test them in the context of my case studies.

I submit that representational interpretation of a vehicle amounts to two tasks: i) establishing the denotative function of the vehicle; and ii) establishing a mapping between vehicle and its target.[44] I proceed to examine these tasks. The first task amounts to deciding the target of a vehicle, namely the object for which the vehicle stands. In other words, a target is a vehicle's putative *denotatum*. A target can be pretty much anything scientists are interested in representing. For example, they cover actual phenomena, possible phenomena, events, concrete objects, states of affairs, just to name a few candidates.

I qualify a vehicle's *denotatum* as "putative" to account for a widely discussed case in the literature on representation, namely that of representation of fictional entities. A vehicle can be used to represent a fictional target. However, in such case, the vehicle does not denote the fictional target, in the sense that there is no actual thing to be denoted. For example, Maxwell's famous diagram of the ether is used to represent the ether. However, there is no (real) ether to be denoted. Thus, representation does not require actual denotation. Nonetheless, in practice, a vehicle that is used as a representation has a "denotative function" (see Suarez, 2015: 44-5). As Elgin (2009) points out: "To be a representation, a symbol need not itself denote, but it needs to be the sort of symbol that denotes" (78). Thus, establishing the denotative function of a vehicle amounts to deciding its target, whether the latter is real or fictional, concrete, or abstract.

The second task – i.e., mapping – amounts to stipulating correspondence relations between parts of the vehicle and parts of the target. This task requires a previous non-representational interpretation of the vehicle, which provides it with a conceptualization and schematization, establishing its component parts. A parallel non-representational interpretation must be conducted for the target as well. (For matters of scope, I do not examine the conceptualization and schematization of the target. I recognize that it is performed in scientific practice, while its workings remain a "black box" in my argument.) Once the schematizations are established, correspondence between component parts of the vehicle and component parts of the target can be established. In this sense, mappings can be

[44] These two tasks are comparable to Hughes's (1997) "denotation" and "interpretation", respectively.

characterized as a fine-grained denotational scheme between component parts in the vehicle and component parts in the target.

As an insightful illustration, consider a kind of mapping discussed by Contessa (2007), which he calls "analytic interpretation." This mapping involves: i) a non-representational interpretation of the vehicle and target in the form of a schematization (i.e., identification of parts, properties and relations); ii) a representational interpretation in the form of a denotational scheme between the schematization of the vehicle and that of the target. In Contessa's words: "An analytic interpretation of a vehicle in terms of the target identifies a (nonempty) set of relevant objects in the vehicle ($\Omega^V = \{o_1^V,\ldots o_n^V\}$) and a (nonempty) set of relevant objects in the target ($\Omega^T = \{o_1^T,\ldots o_n^T\}$), a (possibly empty) set of relevant properties of and relations among objects in the vehicle ($P^V = \{{}^nR_1^V,\ldots {}^nR_m^V\}$ where ${}^nR$ denotes an *n*-ary relation and properties are construed as 1-ary relations) and a set of relevant properties and relations among objects in the target $P^T = \{{}^nR_1^T,\ldots {}^nR_m^T\}$, and a set of relevant functions from $(\Omega^V)^n$ – that is, the Cartesian product of Ω^V by itself *n* times – to Ω^V ($\Phi^V = \{{}^nF_1^V,\ldots {}^nF_m^V\}$, where ${}^nF$ denotes an *n*-ary function) and a set of relevant functions $(\Omega^T)^n$ to Ω^T ($\Phi^T = \{{}^nF_1^T,\ldots {}^nF_m^T\}$)" (57).

As a special case, there are mappings in which the component parts of the vehicle are not mapped to component parts of the target but imputed to the target. That is, instead of establishing a bijective function between parts in the vehicle and parts in the target, the parts in the vehicles are taken to be the parts in the target. This is a common strategy in cases in which the target phenomenon is a black box. In this case, a partitioning of the target is not straightforward. Scientists' best guess is to attribute to the target the schematization of the model.

There are various ways in which mappings can be realized. Consider the various morphisms discussed in the literature. Broadly speaking, morphisms are mappings based on mathematical or structural correspondence between vehicle and target. For example, philosophers of science have studied the virtues of isomorphism (e.g., van Fraassen, 2008), partial isomorphism (e.g., French, 2003), and homomorphism (e.g., Bartels, 2006) as mappings. Other mappings focus on the resemblance between vehicle and target in more

qualitative terms. For example, Weisberg's approach to similarity (2013) explicitly takes into consideration the role of attributes in similarity. Also, Giere (1988) argues that similarity obtains between model and target but in limited respects and only to a limited degree of accuracy (93; see also "model view" in Teller, 2001: 396).

Although I submit that representational interpretation requires mappings, I do not advocate a particular approach to mapping. As Suarez (2004) argues, the specific kind of mapping in the context of scientific representation varies from case to case (776). This position contrasts with the various philosophers who argue that one or another approach to mapping is the right constituent of an account of representation. In more technical terms, my approach can be called "deflationary," i.e., an approach that does not specify the necessary and sufficient conditions for representation (in this case, in terms of specific mappings). This deflationary approach is lucidly expressed by van Fraassen (2008): "There is no representation except in the sense that some things are used, made, or taken, to represent some things as thus or so" (23).

Deflationary approaches are not intended to render discussions on representation as idle. They only submit that such discussions are not to be shaped in terms of specifying necessary and sufficient conditions. For example, van Fraassen (2008) and Suarez (2015) adopt deflationary approaches to scientific representation, but they do engage in discussions on representation in specific terms. In particular, they advocate a used-based approach to representation with models. That is, their discussions are shaped by looking at actual representational practices of scientists. I follow this line of work and refer to this deflationary approach as "pragmatic" approach.

In this pragmatic approach, mappings are an expression of the representational usage of vehicles as models of a target. A mapping affords representation insofar as the mapping allows scientists to use a vehicle as a scientific model that represents a target. In addition, a mapping affords a better or worst representation only relative to the achievement of those practical motivations of scientists using the vehicle as a model. Pragmatic considerations also shape the first task of representational interpretation, namely establishing a denotative function (i.e., deciding the target). That is, an object becomes a target of a model only for a

user with certain motivations. One or another target can be better or worse depending on the scientists' motivations.

There is a wide variety of motivations that prompt scientists to use vehicles as models that represent a target. However, as a general one, I submit that scientists use vehicles as models in a representational mode to engage in "surrogative reasoning" (cf. Contessa, 2007; Suarez, 2004). That is, scientists use vehicles as a surrogate system in which they can reason about the target and generate inferences. By inspecting and manipulating exemplified features in a vehicle, scientists learn facts about the vehicle in question. Subsequently, these lessons can be imputed to the target in conformity with the representational interpretation (i.e., denotational function and mapping to the target). In this sense, mappings are rules of inference between vehicle and target. This approach to representation is often referred to as the "inferential conception" of representation. I discuss this topic further in the next chapter.

To test my ideas on representational interpretation, I take a look at the OFC model. The representational interpretation embodied in the OFC model is somehow hybrid. I suggest that there are two targets in the OFC model, namely BK's spring-block system and seismic earthquakes. On the one hand, OFC claim that their model "is directly *mapped* into a two-dimensional version of the famous Burridge-Knopoff spring-block model for earthquakes" (1244; my emphasis). In fact, OFC provide a sketch of a two-dimensional spring-block system, which they use as a basis to design their cellular automaton. This is explicitly stated: "For the purpose of *mapping* the spring-block model into a cellular automaton model we define an L x L array of blocks by (i,j) […]" (1244; my emphasis). Furthermore, the algorithm in the OFC model is purposely designed to be mapped to the spring-block model: "Thus, the *mapping* of the spring-block model into a continuous, nonconservative cellular automaton modeling earthquakes is described by the following algorithm […]" (1245; my emphasis). That is, the OFC vehicle is representationally interpreted as a model of BK's spring-block system. In simple words, the OFC model is a model of a model.

On the other hand, the OFC model is also intended to represent seismic faults and earthquake occurrence. The intention of modeling earthquakes with their model is stated repeatedly throughout OFC's paper, from its title to the conclusions. Furthermore, the OFC

model is not just a representation of earthquake faults, in the sense of merely standing for them. Rather, it is purposely designed as a surrogate system to study earthquakes in seismic faults. For example, OFC claim that their model allows them "to explain the wide variances in the Gutenberg-Richter law […]" (1244). The fact that aspects of earthquakes are explained by means of investigating the OFC vehicle implies that the OFC vehicle is used as a surrogate system. Thus, I suggest that the OFC model has two targets – i.e., two denotata belonging to two different representational interpretations – namely i) BK's spring-block system, and ii) earthquake faults.

These two representational interpretations are not in conflict, but their compatibility can be differently construed. On the one hand, there is a reading by which one interpretation mediates the other. More explicitly, I suggest that, by representing the BK model, the OFC model also represents earthquake faults due to a transitivity of representation. That is, if the BK model represents earthquake faults, and the OFC model represents the BK model, then the OFC model also represents earthquakes faults. This transitivity is by no means necessary. Not all models that represent another model represent also the target of the latter. The key is that the transitivity of the *denotata* is intentional. More explicitly, OFC intend to study earthquakes by studying an old model of earthquakes, namely the BK model. This form of transitivity of targets is clearly expressed in a context very similar to that of OFC's research: Bak and Tang (1989)'s cellular-automaton model of earthquakes (a precursor of the OFC model). They claim that their model "is actually very close to the generally accepted "block spring" picture of earthquakes […] This is precisely why we believe that our results apply to earthquakes" (15636).

On the other hand, there is a second reading by which the two representational interpretations – the BK model as target and earthquakes faults as target – are independent. In this sense, the OFC model is not only a surrogate system, studied to reason about earthquakes. It is also a "substitute system" which is studied for its own sake. Mäki (2009b) characterizes substitute systems as "freely floating subject[s] of inquiry, unconstrained by any concern as to how [they] might be connected to real-world facts" (36). In this sense, substitute systems are often designed or chosen to test principles or explore its dynamics, without concern for their representational capacities.

The vehicle of the BK model – i.e., the spring-block system – has become a hybrid system in this sense. It is clearly a surrogate system for reasoning about earthquake faults in the context of BK's research. However, it has also become an object of study in itself. This is in part reflected in passages of BK's paper. But more importantly, it is attested by the large number of papers published on the dynamics of spring-block systems *as* spring-block system. A quick search in Google Scholar shows papers of this sort. For example: "Dynamic phases in a spring-block system" (Johansen *et al.*, 1993), "Chaos in a simple spring-block system" (de Sousa, 1995), "Regimes of frictional sliding of a spring–block system" (Putelat *et al.* 2010), to name a few. Naturally, the results of studying spring-block systems can be used to learn about earthquakes if the appropriate representational interpretation is conducted. In addition, studies on the dynamics of spring-block systems can be representationally interpreted to surrogatively reason about targets other than earthquake faults. Indeed, spring-block systems have been used as models of dissimilar targets, from biomaterials (Costagliola *et al.*, 2017) to highway traffic (Járai-Szabo *et al.*, 2010). Because the spring-block system can be taken as a substitute system, the fact that the OFC model represents it does not imply representation of seismic faults. In sum, spring-blocks systems and seismic faults are, as targets, independent of each other. However, when the spring-block system is taken as a surrogate system for reasoning about earthquakes, representation of the spring-block system brings with it a transitive representation of earthquake faults.

I consider it enlightening to introduce a model proposed by Marcelo Dascal (2003) to account for the situation described above with the OFC model and its two targets. Dascal provides a model – called the "onion model" – to account for the meaning or significance of an utterance. The model posits a "hierarchical structure of 'layers of meaning,' which jointly and differentially contribute to the overall significance of a conversational utterance, to what it actually conveys in a particular context" (154). It is interesting to note that, as part of his model, he introduces "core" and "outer shell" layers that contribute to the overall significance of an utterance (ibid).

Extrapolating Dascal's main ideas to the context of scientific models, I suggest that a vehicle's meaning can be fruitfully analyzed in light of the onion model. As my case study

shows, a vehicle might have multiple targets. These targets can be conceptualized as 'layers of meaning'. The various targets jointly contribute to the overall significance of the vehicle. However, a particular target can contribute more or less to the overall significance of the vehicle in a particular context. In other words, there is a differential contribution of the target to the overall significance of a vehicle. This statement recalls my earlier discussion on core and peripheral commitments and their differential contribution to a research program. This is no coincidence. I suggest that scientists can be more or less committed to representational practices, including the denotative function of a vehicle. Following Dascal, the degree of commitment to a target – as contributing more or less to the overall significance of a vehicle – is a matter of context.

A similar thesis is presented by Mäki (2009b). In discussing Schelling's model of racial segregation, he argues that the model involves a "hierarchy of representations: a representation (say words and figures in Schelling's *Micromotives and Macrobehavior*) of a representation (checkerboard on Schelling's desktop) of a representation (imagined city) of a real target system (real-world cities)" (35). Mäki goes further and establishes a link between the reading of "independent layers of targets" and that of "transitivity between targets": "Given this sort of hierarchy, it is also possible to cut across it by talking about the checkerboard (as a concrete object) as a model of real cities (as concrete objects). This would be so by virtue of the supposition that representation is transitive: if A represents B, and B represents C, then A represents C" (35).

In sum, the OFC model is representationally interpreted as being a model of the BK model and of seismic faults. Both targets contribute to the OFC's overall significance as a representation. However, in particular contexts, one or the other target plays a major role. For example, scientists interested in the dynamics of earthquakes would likely assign a major role to the target of seismic faults in the overall significance of the OFC model. In contrast, scientists interested in the dynamics of spring-block systems, even for applications beyond seismology (e.g., biomaterials design), would likely assign a minor role to seismic faults as target of the OFC model. In other words, scientists' commitments play a role in deciding how a vehicle is interpreted. In the following section I develop this idea.

5.5 Interpretations are Committal

In using a vehicle as a model, scientists must provide the vehicle with an interpretation. But this interpretive task can be conducted in a plurality of ways. For example, in the case of non-representational interpretation, a scientist could choose one or another concept to conceptualize the vehicle. Or she could partition the vehicle in different schemes, using different component parts. Or she could emphasize different sets of properties for the vehicle to exemplify. Similarly, representational interpretation can be variously conducted. A scientist could use the vehicle to represent a variety of targets. And, she could choose one or some of the multiple ways available to map the vehicle to these targets. Given this wide range of options, it is worth discussing how interpretations are to be conducted in practice. As I stated above, I advocate a pragmatic approach to interpretation. That is, the interpretation of vehicles depends on the user's motivations in employing the vehicle as a model. These motivations are part of the user's scientific commitments. It is in this sense that interpretations are committal.[45]

In the following, I focus on the role of one kind of commitment in shaping interpretations, namely that of explanatory commitments. My focus on explanatory commitments responds to my interest in the role of models in the crafting of model explanations. Because of this, I spend the next few paragraphs introducing the notion of model explanation before I tackle the issue of explanatory commitments in non-representational interpretation. To be clear, the next paragraphs should be taken as a first approach to model explanation. I continue discussing this topic in the next chapter in further depth.

5.5.1 Model Explanations

I use the term 'model explanation' to refer to explanations whose explanantia make reference to a scientific model or features of a scientific model. This is a minimal feature of model explanations, which is also addressed by Bokulich (2011) in her influential account on

[45] The literature on "perspectivism" provides further support to the thesis that the interpretation of vehicles is committal, e.g., see Massimi (2018), van Fraassen (2008), and Giere (2006).

the subject. However, she includes two additional features in her account of model explanation: i) the counterfactual structure of the model must be isomorphic to the counterfactual structure of the target; and ii) a justification step that specifies the domain of application of the model and shows that the target falls within the domain (39). Given that Bokulich's account of model explanation is intended to be a general account of model explanation (38), I object these two additional features. I submit that they already embed specific commitments. This is a problem, because I consider that a general account of model explanation should be open to admitting different commitments. I proceed to present the specific commitments that these two additional aspects embed.

First, Bokulich recognizes the influence of Woodward's (2003) manipulationist account of explanation in her characterization of model explanations. Commitment to the manipulationist account is expressed in the demand of isomorphism between the counterfactual structures of model and target. To be sure, this might be a common desideratum in several explanatory enterprises. However, it is not the only mapping that can afford explanations. Other morphisms, and even a broader similarity approach, can be so used, depending on the desiderata of the scientists. Furthermore, considering that model explanations can provide plausible or even possible explanations, isomorphism between the structure of the vehicle and target seems too demanding. To be fair, Bokulich's account aims for genuine explanations (38). This roughly means that, for Bokulich, model explanations should correctly capture real patterns of structural dependencies in the world (45). But then, her account can hardly be characterized as a general account of model explanation, excluding modal explanations with models (e.g., how-plausibly and how-possibly explanations, see Chapter 6).

Second, Bokulich argues that, in using a model for model explanations, a domain of applicability for the model should be specified, and it should be shown that the target explanandum falls within such domain. This is referred to as a justificatory step: because the target explanandum falls within the domain of application of the model, a scientist is justified in using the model for model explanations. While I do consider that a justification strategy should be available, I reject Bokulich's articulation of this condition. Her approach suggests that a scientist should provide theoretical and/or empirical reasons that make the model a

"good model" for explaining a target phenomenon (39). I consider this too strict. In my view, a scientist is justified in using a model for model explanations simply because the scientist's commitments make it acceptable. And a scientist's commitments may be justified epistemically but also practically, making Bokulich's strictly theoretical and empirical reasons too narrow (see Chapter 3).

In a later paper, Bokulich (2012) elaborates on her justificatory step and reaches a conclusion more in line with my views. This time, she attempts to distinguish fictional models that are explanatory from those that are not. She argues that scientists are justified in using fictional models to explain phenomena if the model in question is "adequate." She argues that what counts as an adequate fictional model is "negotiated by the relevant scientific community and will depend on the details of the particular science, the nature of the target system, and the purposes for which the scientists are deploying the model" (734). I consider this more contextual articulation as an improvement. Unfortunately, by the end of her paper, Bokulich allows scientific realism to play a central role in deciding what is an adequate model for model explanation.

In the way I construe them, model explanations are not readily available in a model. Instead, I suggest that a model which embodies a certain interpretation may have resources that are exploitable for explanatory purposes. In this sense, model explanations must be put together by the user of the model. Jebeile and Graham Kennedy (2015) present a similar approach, in which model explanations are posed as an activity. In this sense, the tasks involved in model interpretation are extended to the explanatory domain. The success of explanatory enterprises with models depends upon features of the model – as afforded by the interpretation – and features of the user of the model. On the one hand, the model must be the sort of thing that allows the extraction of explanatory propositions. On the other hand, the user of the model must be a skilled scientist who knows how to extract explanatory propositions from the model. In other words, a model is not intrinsically explanatory. The

explanatory propositions – which are afforded by the model – must be put together by an explainer (ibid: 387).[46]

The format of model explanations may vary widely. In this sense, model explanations may be classified in terms of different accounts of explanation. For example, model explanations may be classified as mechanistic, mathematical, unificationist, and so on. The account of explanation that a particular model explanation instantiates depends on how the vehicle in question is interpreted. Different interpretations of the vehicle enable different kinds of model explanations, or so I argue. My thesis is that explanatory commitments guide the non-representational interpretation of the vehicle, and this affords specific explanatory features to the resulting model. There is a qualification to this thesis. I conjecture that the influence of explanatory commitments is manifested in both modes of interpretation, i.e., non-representational and representational interpretation. However, for the purposes of this dissertation, I focus on studying the influence of explanatory commitments in guiding only the non-representational component of the interpretation of vehicles.[47]

5.5.2 The Role of Explanatory Commitments in Non-Representational Interpretation

I focus on non-representational interpretation for the following reason: non-representational interpretation of a vehicle involves the tasks of conceptualization, schematization, and exemplification. These tasks are the middle link between explanatory commitments and the resulting interpretation upon which model explanations are based. In other words, conceptualization, schematization, and exemplification constrain the potential kinds of model explanations that can be derived from the resulting model. This is because distinct kinds of explanation focus on particular concepts and structures.

[46] Rohwer and Rice (2016) present a more nuanced argument, in which various relations between models and explanations may exist. In particular, they argue that models can be regarded as intrinsically explanatory because they can be identified with explanations. However, in order to make this claim, Rohwer and Rice assume that models can be characterized as sets of propositions. This is a strong claim. Not all models are propositional. Still, Rohwer and Rice suggest that it suffices for models to be *reinterpretable* as sets of propositions.

[47] Rice *et al.* (2018) discuss "explanatory interpretation" as a crucial step in explaining with models (14). However, their treatment of this notion is representational. In their discussion, the models embody an established non-representational interpretation, namely a mathematical-statistical one. Explanatory interpretation, in their view, amounts to the way in which the consequences of these mathematical models are used to explain real-world target phenomena.

In order to assess the impact of explanatory commitments on the non-representational interpretation of vehicles, I assume a methodological simplification. I assume that explanatory commitments are clustered in well-established accounts of explanation. It is important to recall that, in Chapter 4, I actually showed the opposite. In this sense, this assumption is adopted only for matters of convenience. As far as I have reflected on this assumption, I consider that it does not make an argumentative contribution to my thesis besides its economy. In fact, I suggest that my argument can be generalized to any coherent set of explanatory commitments beyond well-established accounts of explanation.

Having clarified the scope of my assessment, I proceed to analyze the impact of explanatory commitments upon non-representational interpretation. First, consider the task of conceptualization of a vehicle. Scientists collaborating within research programs are typically committed to a common conceptual framework. These conceptual frameworks provide candidate concepts to represent the vehicle. Explanatory commitments do not only contribute to conceptual frameworks with their own concepts. They also allow to navigate an accepted conceptual framework and choose concepts that are more or less suitable to fulfil the explanatory desiderata. For example, consider concepts such as 'mechanism', 'process', 'variable', 'graph', 'conserved quantity', and 'cellular automaton'. These are distinct concepts that are typically associated with distinct accounts of explanation. Commitment to one account of explanation or the other prompts choosing specific concepts to conceptualize a vehicle.

As an illustration, consider the OFC model. And consider explanatory commitments of a mathematical and new mechanistic sort. My suggestion is that, in light of these distinct commitments, scientists are guided to conceptualize the OFC vehicle differently. For example, scientists committed to mathematical explanations might aim to conceptualize the OFC vehicle as a graph, a cellular automaton, a network, or another suitable abstract mathematical notion. In contrast, scientists committed to mechanistic explanations are prone to conceptualize the OFC vehicle as a system, a mechanism, a machine, and the like. As a matter of fact, OFC conceptualize the OFC vehicle as a cellular automaton. This is clearly stated in the very title of their paper: "Self-Organized Criticality in a Continuous,

Nonconservative *Cellular Automaton* Modeling Earthquakes" (1244: my emphasis). This suggests that OFC are committed to explaining aspects of the behavior of earthquakes at an abstract level, mathematically.

Certainly, other (i.e., non-mathematical) explanatory commitments can be put to use in guiding alternative conceptualizations of the OFC vehicle. As an illustration, consider similar research conducted by Bak and Chen (1989). In their paper, Bak and Chen discuss a cellular automaton that simulates the behavior of earthquakes. The following passage reflects the mathematical approach to the non-representational interpretation of the vehicle:

> *"A discrete variable Z = 0, 1, 2, 3... is defined on a d-dimensional lattice. The variable at a particular site and its 2d nearest neighbors are updated according to the simple rule*
> $Z \rightarrow Z - 4;\ Z_{nn} \rightarrow Z_{nn} - 1 \quad if\ Z > Z_{cr}$,
> *and it is driven by letting*
> $Z \rightarrow Z + 1$,
> *at some random position" (6).*

However, in a later paper, Bak and Chen (1991) discuss their cellular automaton model with a different approach. Although Bak & Chen explicitly state that their model is a computer model, they depict it as a spring-block system (see Figure 5-1). The figure is accompanied by the following caption: "EARTHQUAKE MODEL simulates the forces on blocks of the earth's crust. Whenever the force on a block exceeds a critical value, the block slides, and the force is transferred to neighboring blocks. Each white square represents sliding blocks; each cluster represents an earthquake […]" (49). This caption is striking: Bak and Chen (1991) describe their model in terms of interacting blocks, while Bak and Chen (1989) describe their model mathematically as a cellular automaton.

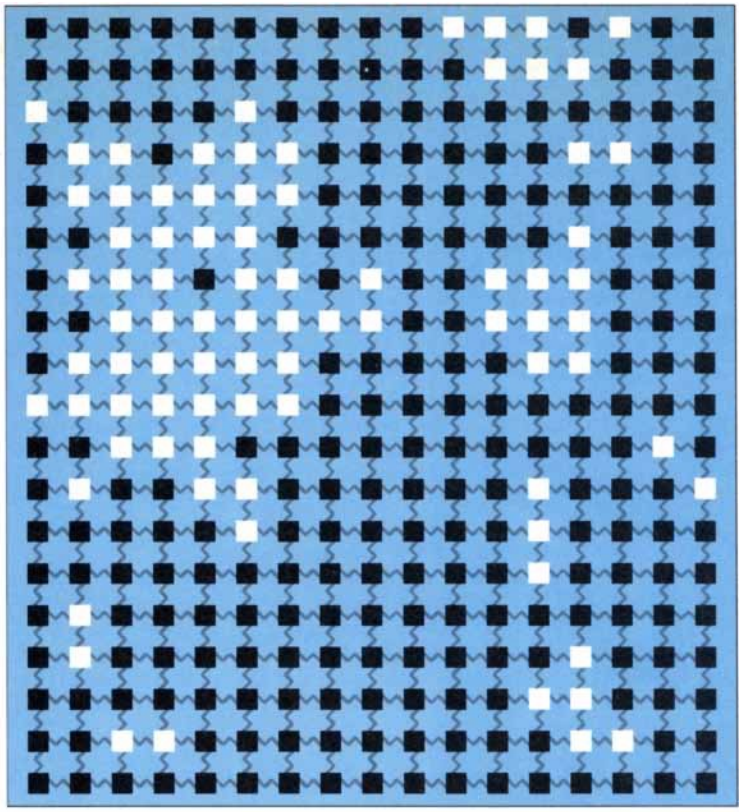


Figure 5-1: Earthquake model (Bak & Chen, 1991: 49).

One explanation for this difference is that Bak and Chen are simply using terminology associated with a target system – namely a spring-block system – to expediently talk about their cellular automaton. It is simply figurative speech. In this sense, there is no conflict: their model is in fact conceptualized as a cellular automaton which represents a spring-block system. However, another hypothesis is that there is an alternative conceptualization of the vehicle in place. That is, Bak and Chen are not simply saying that their cellular automaton represents blocks and springs. Rather, Bak and Chen construe the vehicle as a spring-block system. This case shows that what was once conceptualized as a cellular automaton in a given context can be conceptualized as a spring-block system in another context.

Depending on the accepted explanatory commitments, scientists might prefer one or another conceptualization. This is because conceptualizing a vehicle as – for example – a mathematical entity or as a concrete mechanism has consequences on the kind of explanations that can be derived from the respective models. As I argued in the previous chapter, there might be an overlap of commitments in attempting to explain mathematically

or mechanistically. In my case studies, BK and OFC seem to overlap in their commitment to explanations based on discreteness and modularity (see Chapter 4). But these specific commitments are ultimately embedded in distinct broader explanatory commitments. As I argued above, BK's commitments are better classified as mechanistic while OFC's commitments are more adequately classified as mathematical.

Once the conceptualization is in place, the schematization of the vehicle is based on the inner structure of the employed concepts. This inner structure might be more or less well-defined within a certain program. For example, the concept of 'graph' has a rather stable definition across mathematics. In contrast, a concept such as 'mechanism' has been used with different senses across scientific disciplines. Furthermore, within the philosophy of science, the notion of 'mechanism' has been used in a wide variety of ways (cf. Illari & Williamson, 2012). While a consensus might be desirable, it is not necessary for schematization: a concept might have different inner structures across research programs. As long as some inner structure is specified – in line with the commitments held within program – a schematization can be conducted.

As an illustration, consider two distinct explanatory concepts, namely 'graph' and 'mechanism'. A graph is a collection of nodes and edges. Thus, the inner structure of 'graph' – i.e., its subconcepts – involves 'nodes' and 'edges'. Accordingly, the schematization of the OFC vehicle amounts to mapping (or labelling) parts of the OFC vehicle as 'nodes' and 'edges'. A straightforward schematization is to take each cell of the cellular automaton as a node, and the relations of contiguity between neighboring cells as edges between nodes.

Now, take the concept of 'mechanism'. Following Illari and Williamson's (2012) approach, a mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon (120). This account of mechanism is intended to capture various senses of mechanism scattered in the literature, i.e., it is intended as a minimal account. For this reason, it is a convenient candidate explication of the concept of 'mechanism' to retrieve an inner structure. According to Illari and Williamson's account, the elements of 'mechanism' are 'entities', 'activities', 'organization' and 'phenomenon'. These elements can be taken as the subconcepts of the concept of 'mechanism'.

A schematization of the OFC vehicle as a mechanism amounts to labelling parts of the OFC vehicle as 'entities', 'activities', 'organization', and 'phenomenon'. First, consider mechanistic entities. Entities are the parts of a mechanism that engage in activities and are spatiotemporally located (MDC, 2000: 3, 5). In this sense, cells are a good candidate for mechanistic entities. Cells engage in activities – if we conceptualize mathematical relations as activities. And cells are spatiotemporally located – if we conceptualize the mathematical relations of neighborhood as a virtual spatiotemporal location. Furthermore, cells are fundamental units: they are part of a cellular automaton, but they have no proper parts. In particular, they are not part of each other and there are no overlaps among them.

Second, mechanistic activities are the "producers of change" in a mechanism (MDC, 2000: 3). I suggest that driving and relaxation produce all changes in the OFC model. Indeed, changes in the OFC model are limited to modifications in the states of cells. And the states of cells change only due to driving and relaxation instructions. Thus, driving and relaxation are the activities in the OFC model.

Third, organization is "whatever relations between the entities and activities discovered [that] produce the phenomenon of interest" (Illari & Williamson, 2012: 128). Organizational features in the OFC vehicle are various. For instance, the (virtual) spatiotemporal relations between cells can be taken as a mechanistic organizational feature. Or, to take a different example, the temporal sequencing of activities (slow driving and fast relaxation) is another mechanistic organizational feature. Finally, if we take a phenomenon to be "something like [a] 'characteristic activity'" (Illari & Williamson, 2012: 124), then avalanches are a natural candidate to be conceptualized as the mechanistic phenomenon in the OFC vehicle.

Finally, consider the issue of exemplification. Depending on scientists' explanatory commitments, one or another feature of the vehicle might be more or less suitable to highlight. For example, consider the OFC vehicle conceptualized and schematized as a mechanism. Scientists using the OFC vehicle as a model might want to draw attention to a particular aspect of the vehicle interpreted as a mechanism, say a subset of its activities. This

decision is based on the explanatory interests of scientists. In this case, the activities are exemplified because they play an explanatory role, either as part of the explanandum or explanans.

In contrast, if the OFC vehicle is conceptualized as a graph, a scientist might highlight some of its topological properties. For instance, the OFC vehicle exemplifies being a regular lattice.[48] That is, the number of edges a node shares with other nodes are the same across the bulk of the cellular automaton. As another topological property, the OFC vehicle exemplifies non-modularity. That is, there are no sets of nodes that share more edges with each other than with the rest of the network. In other words, the OFC model has no community structure or clusters.

As I argued by the end of last chapter, BK's explanatory commitments can be adequately classified as mechanistic, while OFC's are better classified as mathematical. These different commitments guide the non-representational interpretation of the respective vehicles. Accordingly, the BK model, as interpreted by BK, prompts model explanations of a mechanistic kind. And the OFC model, as interpreted by OFC, foster model explanations of a mathematical kind.

These conclusions must be qualified. First, this is assuming that BK and OFC's commitments are adequately classified as mechanistic and mathematical, respectively. I argued for this in the last chapter. I expect I made a convincing case. Second, the BK and OFC may be interpreted differently. This is particularly important in the OFC case, where the conceptualization is somehow hybrid between mechanistic and mathematical. In the end, I consider that a mathematical treatment has priority, given the design of the vehicle as a cellular automaton.

This way, I conclude this chapter. At this point, it is convenient to take stock. I have argued that scientific models can be characterized as interpreted vehicles used for scientific purposes. One such purpose is the attainment of explanations of phenomena based on the

[48] Caruso *et al.* (2006) explore the robustness of SOC in the OFC model by testing different topologies.

resources of the models in question. These explanations are referred to as "model explanations." In order to be used as models, vehicles must be interpreted, minimally in a non-representational mode, but usually in tandem with a representational interpretation that relates the model to a target. These interpretative tasks are shaped by scientific commitments. In particular, I have argued that the non-representational interpretation of vehicles is shaped by explanatory commitments in a research program. Given that non-representational interpretation of vehicles is influenced by explanatory commitments, the resulting model contains features that are exploitable for explanatory purposes. In particular, the model is conceptualized and schematized in ways that provide the resources for explanation. Furthermore, the model can be taken to exemplify specific features, which are the main resources for model explanations. Just to be clear, the resulting model is not intrinsically explanatory. Rather, the user of the model must be a skilled practitioner who is able to extract the relevant explanatory information from the model to put together an explanation. In the next chapter, I explore the epistemic status of these resulting model explanations.

5.6 Précis

The main theoretical theses of this chapter are:

- Thesis 1: Interpretation embodies an activity-output duality: it is an activity of bestowing meaning, and it is the outcome of such activity.
- Thesis 2: Scientific models are vehicles interpreted for scientific purposes. As a consequence, a same vehicle interpreted differently embodies different models.
- Thesis 3: Interpretation of vehicles has two components. Minimally, it involves a non-representational interpretation. In addition, it may involve a representational interpretation.
- Thesis 4: Non-representational interpretation of a vehicle involves conceptualizing the vehicle, schematizing it, and deciding the features that is exemplifies.
- Thesis 5: Representational interpretation involves the tasks of deciding the target of the resulting model and the mapping between vehicle and target.

- Thesis 6: In its two modes, interpretation is guided by commitments. In particular, if scientists aim for model explanations, explanatory commitments guide the non-representational interpretation of vehicles.
- Thesis 7: Non-representational interpretation constrains the kinds of model explanations that can be delivered in a model, by means of establishing the relevant concepts and structures.

The main lessons learnt from my case studies are:

- BK's mechanistic explanatory commitments prompt a non-representational interpretation of the BK vehicles as a spring-block system.
- OFC's mathematical explanatory commitments prompt a non-representational interpretation of the OFC vehicle as a cellular automaton.
- Both the BK and OFC models are representationally interpreted as models of earthquakes. In addition, the OFC model is also representationally interpreted as a model of the BK model.

***

6 Genuine Model Explanations

In the previous chapter, I introduced the notion of model explanation and claimed that a model explanation is an explanation whose explanans makes reference to the content of a scientific model. But I said little about the conditions that make a model explanation genuinely explanatory. The purpose of this chapter is to close that gap. I suggest that the epistemic status of model explanations depends on the resemblance between model and target. There are various nuances to this claim, which I discuss throughout this chapter. A main problem with this claim is that scientists often explain target phenomena by using idealized models which do not resemble their targets. I argue that model explanations based on idealized models may still be genuinely explanatory, even if they are not actually true. I submit that this is admissible provided that model explanations are qualified with a modal term.

The structure of this chapter is as follows. In section 6.1, I introduce the topic of surrogative reasoning and argue that model explanations are a special case of surrogative reasoning. I also establish the minimal conditions for surrogative reasoning. In section 6.2, I proceed to analyze the epistemic status of model explanations. I begin with an examination of the minimal status of model explanations, namely that of potential explanation. Subsequently, I discuss the challenges of pitching a higher epistemic status for model explanations based on idealized models. In section 6.3, I sketch a solution to these challenges. I first engage in a discussion on modalities of explanation. This discussion provides me with the resources to deliver an account of model explanations based on an idealized model in which they could be imputed to a target as genuinely explanatory. I begin now with my discussion on surrogative reasoning.

6.1 Preliminaries on Surrogative Reasoning

Surrogative reasoning is the process of drawing inferences about a system of interest by means of studying another system (Swoyer, 1991). These other systems are referred to as

"surrogate systems." Given that surrogate systems are used to reason about systems of interest, the former are representations of the latter – assuming an inferential conception of representation (Suarez, 2004). Scientists often resort to surrogate systems because the systems of interest – hereinafter, the targets – might be difficult to access, too complex to understand, or unsuitable for manipulation. In contrast, surrogate systems are typically selected or constructed in a way that capitalizes on their being accessible, intelligible, and manipulatable.

Scientific models are commonly used as surrogate systems. There are various reasons for this. Typically, scientific models are objects to which scientists have direct access. This allows scientists to investigate the models and acquire insights about their inner workings via observation and manipulation. Furthermore, scientists are often involved in the design and construction of models. This provides the scientists with control and knowledge about the model's architecture. Due to these qualities – accessibility and knowability – scientists are in an advantageous position to investigate models and learn about them. Models – as surrogate systems – offer the accessibility and knowability that targets withhold. In particular, scientists can investigate the internal dynamics of a model to establish how the resources contained in the model bring about results of interest (cf. "demonstration" in Hughes, 1997: S331-2).

Two aspects of scientific models must be accessible and knowable in order to function as surrogate systems, namely their structure and outputs.[49] Both aspects are relative to a non-representational interpretation of the corresponding vehicle. A model's structure is decided in the schematization of the corresponding vehicle. The notion of structure comprehends the component parts of the model and their relations, whether static or dynamical. A model's outputs are those features exemplified by the model. Given my focus on the BK and OFC case studies, I am particularly interested in those outputs that amount to the results of simulations.

[49] Other elements of models might not be as accessible and knowable as the structure and outputs. For example, it is often the case that the processes taking place in computer simulations are "epistemically opaque" to the scientists (see Humphreys, 2009: 618). This basically means that scientists are not able to survey or even access the computational processes in a simulation (cf. Duran & Formanek, 2018: 649-50). For the purposes of my argument, the opacity of processes in computer simulations is not relevant, *insofar as they are reliable*. If the computer simulations were not reliable, then the inferences drawn from studying them would not be justified. For a discussion on the sources of computational reliabilism and their assessment, see Duran and Formanek (2018: 656-663).

Given that structure and outputs are interpreted with a same set of coherent explanatory commitments, there is potential for inferring dependence relations between them. Furthermore, given that scientists can often manipulate the models, they can often infer counterfactual dependencies between structural features and outputs in a model. As I discuss further below, these dependencies are the basic resources exploitable for explanations of the model's outputs.

Naturally, surrogative reasoning is not complete without relating the lessons learnt in the context of the model to the target. Surrogative reasoning is, first and foremost, about "imputation," i.e., about ascribing the acquired knowledge of the model to the target. While most of the investigation of the model can be conducted non-representationally, the imputation of the lessons to the target requires a representational interpretation. That is, a target must be established and a mapping to that target must be in place. This is the core of surrogative reasoning: scientists learn about the model and impute those lessons to the target.

A major concern in using scientific models for surrogative reasoning is assessing the epistemic status of inferences about the target based on the investigation conducted in the model. For the most part, scientists are not interested in drawing just any inference about a target system. Typically, scientists intend to use models as surrogates for the generation of "sound" inferences about the target, i.e., inferences that are true – or approximately true – of the targets (cf. Contessa, 2007: 51). In other words, it is a common desideratum that the representational interpretation of a vehicle results in a "faithful" model, i.e., a model that affords sound inferences about its target (ibid: 54).[50]

A subsequent problem is to decide criteria for the assessment of a model's faithfulness. Various candidates have been proposed, which span different forms of morphisms and qualitative approaches to similarity. I suggest that these candidates can be

[50] The *Internet Encyclopaedia of Philosophy* provides the following characterization of validity and soundness in the context of deductive arguments: "A deductive argument is said to be valid if and only if it takes a form that makes it impossible for the premises to be true and the conclusion nevertheless to be false. Otherwise, a deductive argument is said to be invalid. A deductive argument is sound if and only if it is both valid, and all of its premises are actually true. Otherwise, a deductive argument is unsound."

collectively characterized as different variations of a same general notion, namely that of *resemblance*. The general idea is that models must embody some sort of resemblance to their targets in order to be faithful and thus afford sound inferences. It is this resemblance which justifies imputing the lessons learnt in the context of the model to the target. Mäki (2009b) expresses this idea as follows: "For the acquisition of relevant information and understanding [about the target] to take place, the surrogate system [in this case, a scientific model] must resemble the target system in suitable ways" (37).

Resemblance between model and target may come in various types. For a detailed review of the topic, see Cowling (2017). Regardless of the specific type, one thing is clear: resemblance between model and target – as Mäki says – must be "suitable" for surrogative reasoning. I suggest that this suitability is decided contextually. In particular, the purposes and competences of the user of a model constrain the respects and degrees in which the model must resemble its target. In other words, I advocate a pragmatic approach to resemblance (cf. "relevant similarity" in Parker, 2009: 493; and Roca-Royes, 2017: 233-5).[51]

As an illustration of the pragmatic approach to resemblance, consider an updated map of the London Underground (I adapt this example from Contessa, 2007). The map is a faithful representation *for the purposes of navigation for a competent commuter*. But it is not straightforwardly a faithful representation for other purposes of just any commuter. For instance, if a commuter intends to infer the relative distance between stations, the map is not a faithful representation: the distance among stations in the map is not proportional to the actual distances among stations. Even for the purposes of navigation, the map might not function as a faithful representation for a commuter who lacks some competences to derive inferences from the map, e.g., an illiterate or color-blind commuter.

[51] Admittedly, there are approaches of resemblance that are not based on the user's purposes or competences, i.e., non-pragmatic approaches to resemblance. Instead, these approaches attempt to identify objective relations of resemblance between model and target (see "dyadic approaches" in Knuuttila, 2010: 142). Even Mäki argues that there may be ontic constraints to the issue of resemblance. That is, there are actual aspects about a model and its target that "decide" whether a model resembles its target, regardless of the agent's intentions (see e.g., Mäki, 2009a: 81-2). Having said this, I frame my discussion within a pragmatic approach to resemblance.

In order to assess resemblance between model and target, scientists must be able to compare them. This comparison presumes knowledge about the target, at least in those aspects in which the model is intended to resemble it. The problem is that such knowledge is not always available. In fact, it might well be the case that the very lack of knowledge concerning the target prompts scientists to engage in surrogative reasoning with models. This is a critical conundrum: surrogative reasoning is a practice that aims at learning about a target, but scientists' trust in such practice presumes knowledge about the target.

One way to deal with this conundrum is to admit that some knowledge about the target must be in place. In this sense, surrogative reasoning has the potential to expand on an already existing – even if basic – knowledge of the target. Assuming that the target is a causal phenomenon, I distinguish two types of knowledge that scientists can possess about the target. First, scientists may possess knowledge about the causes of a target phenomenon. I refer to this type of knowledge as "causal" knowledge. Second, scientists may possess knowledge about the phenomenon itself. I refer to this type of knowledge as "phenomenal" knowledge.

As a minimal condition, one of these types of knowledge must be in place in order to assess resemblance with the model. In other words, there must be some tether between model and target for surrogative reasoning. By possessing causal knowledge, the target's causal structure can be compared to the model's structure. And by possessing phenomenal knowledge, the target phenomenon can be compared to the model's outputs. As a corollary, cases in which there is no causal nor phenomenal knowledge of the target are non-starters for surrogative reasoning: there are no resources to assess resemblance.

I proceed to examine two scenarios of surrogative reasoning with different configurations of causal and phenomenal knowledge. I relate these scenarios to typical settings in scientific enterprises, namely prediction and explanation. First, consider the case in which scientists have causal knowledge of the target and intend to gain phenomenal knowledge. This is a prototypical case that calls for predictive efforts: scientists know about causes and their operation in nature and intend to anticipate their eventual effects. In such situations, scientists typically construct models that resemble the causal structure of the

target – in relevant respects, to a certain degree (cf. "representational fidelity criteria" in Weisberg, 2013: 41). I discuss this scenario only briefly, given that my dissertation does not focus on the topic of prediction. The general idea is that a model whose (causal or mathematical) structure resembles the causal structure of the target – in relevant respects to a certain degree – is expected to produce outcomes that resemble the target phenomena. In other words, the model is expected to predict the target phenomena (see Figure 6-1). Certainly, scientists can be more or less inclined to impute the outputs of the model to the target. This depends especially on the extent and quality of the causal knowledge of the target and the scope and degree of similarity achieved by the model's structure.[52]

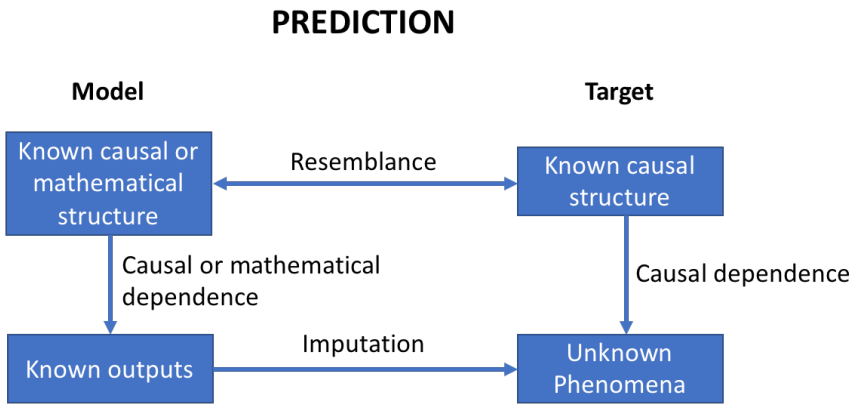


Figure 6-1: Prediction as surrogative reasoning.

In the second scenario, scientists have phenomenal knowledge about the target, but they ignore its causal structure, or intend to learn more about it. These cases typically call for explanatory efforts: scientists know a phenomenon and aim to explain it. The strategy that I address here is that of model explanations: scientists select or build a model whose outputs resemble those of the target. As Rice *et al.* (2018) put it, in order to explain with models, one must show that the explanandum in question "approximates a consequence" of the model (14). This consideration is pragmatically decided (cf. "dynamical fidelity criteria" in Weisberg,

[52] There are models that are used for predictive purposes whose (causal or mathematical) structure does not resemble that of the target, namely phenomenological models. I suggest that scientists are justified in using phenomenological models for predictive purposes based on their actual predictive success. Predictive success can only be attributed to a model if scientists *already have* phenomenal knowledge of the target.

2013: 41).[53] Once the phenomenal resemblance is settled, scientists can focus on examining and learning how the model's structure brings about its outputs. That is, scientists can explain the model's outputs by resorting to the model's structure. The resulting explanation can be imputed to the target minimally as a potential explanation based on the phenomenal resemblance between model and target. It is in this sense that model explanation is a special case of surrogative reasoning (see Figure 6-2).

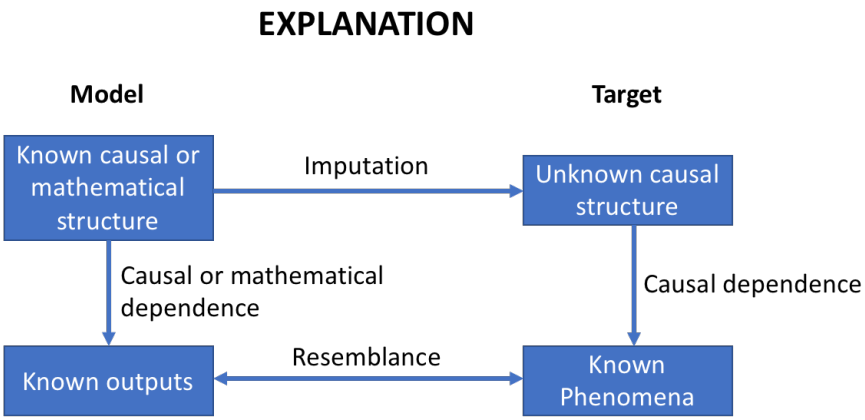


Figure 6-2: Explanation as surrogative reasoning concerning a target's causal structure.

I use the BK model to illustrate some of the basics of surrogative reasoning. The BK model is a surrogate system in the sense of being a system – namely, a spring-block system in the laboratory implementation– used to reason about another system – namely, seismic faults. BK engage in a detailed study of the structure of the spring-block system. They also study thoroughly the outputs of the model (i.e., simulation results). Subsequently, BK compare simulation results to empirical-statistical data of naturally occurring earthquakes. In particular, the results from the laboratory model concerning frequency versus size of shocks and randomness of shocks are explicitly compared to the behavior of naturally occurring earthquakes. This is also the case with simulation results from the mathematical

[53] Sterret (2006) discusses the notions of static mechanical-structural similarity and dynamic similarity (72). These notions are analogous to Weisberg's (2013) notions of representational and dynamical fidelity, respectively.

implementation, especially with regards to the aftershocks sequence. These comparisons suggest that BK use their model as a surrogate system for seismic faults.[54]

However, the epistemic status of the surrogative reasoning conducted with the BK model is *prima facie* unclear. The issue is that the BK model is a "highly simplified" model (342). Because of these simplifications, it is likely that the BK model might miss relevant aspects of the dynamics of earthquake faults. BK mention two main aspects that call for simplification: i) the geometrical configuration of the systems under study (i.e., seismic faults); and ii) the methods of excitation that bring about the earthquakes. Simplifications concerning these aspects are required for practical reasons. As BK claim, simple geometries and excitation methods are "convenient" for the construction of both the laboratory and mathematical implementations (341). Still, this comes at the cost of constructing a model that is unlike the target in various regards, to some degree. These dissimilarities may have an impact on the epistemic status of the resulting surrogative reasoning.

In the following section, I discuss the assessment of the epistemic status of surrogative reasoning in the form of model explanation. In particular, I address some of the challenges that scientists face in this regard. Among these challenges, I focus on model explanations based on models that do not resemble their targets in relevant regards and in significant degrees, namely idealized models.

6.2 Assessing the Epistemic Status of Model Explanations

6.2.1 The Minimal Status: Model Explanations as Potential Explanations

[54] In addition, consider the following suggestive quotes: "We can thus imagine the particles [i.e., discrete entities in which the continuum of rocks at opposite sides of a fault is divided] on opposite sides of the fault reduced to a two-dimensional network of masses interconnected by springs which *represent* the usual elastic elements and which are further coupled by frictional elements interconnecting the elements on opposite sides of the fault" (342; my emphasis). "We require for proper *simulation of naturally occurring earthquakes*, that the rate of drive be sufficiently small that all quake-connected processes such as rupture, radiation and aftershocks be completed while the drive, *represented* in our model by the support block holding the leaf springs, is essentially stationary" (356; my emphasis). "Particles 6-10 [in the mathematical model] form the *analog* of a strongly seismic fault" (359; my emphasis).

The notion of resemblance admits degrees. That is, models and their targets can be more or less similar to each other. Given that resemblance is gradual, an analogous graduality can be attributed to the epistemic status of surrogative reasoning. That is, the lessons learnt in the context of a model are more or less informative of the nature of the target. In particular, I suggest this applies to model explanations. That is, model explanations can be more or less genuinely explanatory about the target, with an explanation being genuinely explanatory if its content is true – or approximately true – about the target explanandum.

The graduality of genuineness in model explanation is not an uncontroversial claim. Traditionally, philosophers of science submit that explanations must be true – or approximately true – in order to be genuinely explanatory. In other words, there is no such thing as a genuine explanation that is false about the target. This tradition can be traced back to Hempel (1965), who demands true premises and laws in order to derive true conclusions in his DN model of explanation. Other contemporary philosophers continue this tradition, such as Kaplan and Craver (2011), Strevens (2008), and Woodward (2003) (cf. Rohwer & Rice, 2016: 1137). I do agree that explanations must provide true content in order to be genuinely explanatory. However, I argue that, even accepting this constraint, model explanations admit a gradual epistemic status. This is the case because there are different modes in which an explanation can be true about the target. In a way, the rest of this chapter can be read as an elaboration of this position.

To assess the epistemic status of a model explanation imputed to a target phenomenon, scientists must be able to assess the resemblance between the explanans of the model explanation and the causal structure of the target. To engage in such assessment, some causal knowledge of the target must be available. Otherwise, there are no resources to evaluate resemblance. Yet, if there is no causal knowledge of the target (or only basic knowledge of it), the model explanation can still be imputed to the target. These are cases in which the target is a "black box," a system whose external behavior can be observed and studied, but whose inner workings are not accessible. Insofar as the model explanation accounts for the outputs of the model, and the outputs of the model resemble the behavior of the target, the model explanation can be imputed to the target as a "potential"

explanation. Given the lack of causal knowledge of the target, the epistemic status of this potential explanation cannot be empirically decided.

I propose a characterization of "potential explanation" along the following lines. Potential explanations are propositions that follow the logic of an account of explanation. In this sense, the adjective 'potential' is a formal notion. No empirical investigation is required to declare an explanation a potential one.[55] This approach is expressed by Strevens (2013) as follows: "A potential explanation, as the term is used in the explanation literature, is one that satisfies what you might call the internal condition for explanatory correctness […] The internal condition holds or fails to hold independently of the way things are in the outside world" (512-3). As a slogan, potential explanations "fit the template" of an account of explanation, resorting to its ontology, structure, and inner logic.[56]

Potential explanations, as their name suggest, have the potential to be correct explanations, i.e., true about an explanandum phenomenon. This potential lies in the validity of their construction, which accords with the logic of an account of explanation. The epistemic status of a potential explanation can be upgraded in light of empirical evidence to a genuine explanation. Still, if the evidence shows that the potential explanation is incorrect, its status as potential explanation remains intact. This is because the potential character of an explanation is formal and thus not modified by refuting evidence.

I proceed to elucidate how model explanations function as potential explanations of the explananda phenomena in the target. I do this in four steps. To begin with, a resemblance relation is established between the explanandum phenomenon in the target and a model's

[55] This approach to potential explanation is inspired by Hempel and Oppenheimer's account of "potential explanans" (1948: part 3). Their definition resorts to elements that are idiosyncratic to their deductive-nomological account of explanation. I prescind from those idiosyncratic elements and retain what I consider the most relevant aspect, namely the relation of derivation between explanans and explanandum (in their case, "deductive" derivation). However, I expand the relation of derivation between explanans and explanandum to a general relation of dependence. This way, I can account for the variety of "logics" of explanation across accounts of explanation.

[56] This "fitness" manifests as some sort of dependence relation between explanandum and explanans. This dependence relation is cashed out differently across distinct accounts of explanation. For example: i) the explanandum could be entailed by the explanans; ii) the explanandum could be more probable given the explanans; iii) the explanandum could be caused by the processes described in the explanans; iv) the explanans posits a unifying pattern to which the explanandum belongs; to name a few cases.

outputs. This phenomenal resemblance is crucial. The idea is that a model cannot be used to explain the target phenomenon, strictly speaking, because the model does not exhibit the target phenomenon. What the model exhibits is a phenomenon that resembles the target phenomenon. It is within the scope of that resemblance that models can be used to explain targets. At this point, the pragmatic nature of resemblance becomes relevant. The resemblance between the explanandum phenomenon in the target and the exemplified outputs in the model is such that it allows both to be characterized as the *same kind of phenomenon*. I do not intend to use the term 'kind' in any strong ontological sense. The model's outputs and the target phenomenon are the same kind of phenomena simply based on agreed respects and up to a certain degree for a user or community of users.[57]

Second, an explanation is crafted for the exemplified outputs of the model. This explanation describes a dependence relation between the explanandum (i.e., exemplified outputs) and the explanans (i.e., features of the model's structure). These dependence relations can be investigated thanks to the accessibility and knowability of the structure and outputs of the model. This is why scientists use models as surrogate systems: because it is more expedient to craft explanations in the context of a model. It is worth noticing that these dependence relations are conceived in alignment with the non-representational interpretation used for the model. In this sense, the explanatory commitments of the model's user are manifest in the crafting of this potential explanation.

Unfortunately, mistakes may be made in judging an explanation of a model's outputs as a correct one. That is, even in the context of models, explanations are fallible. This is not a problem in the context of crafting potential explanations. After all, an explanation is a potential one if it fits the template of an account. However, this becomes a problem if the potential explanation is expected to have a higher epistemic status. In these cases, significant

[57] A similar notion is that of "universality" (see e.g., Batterman, 2002). Broadly speaking, systems that exhibit similar patterns of behavior are said to belong to the same universality class, even if their physical constitution is widely different. I avoid using this notion because universality classes are intended to refer to objective entities, which can be investigated via empirical research. My notion of "same type of phenomena" is stipulative and it responds to pragmatic considerations. Admittedly, the boundary between objective and pragmatic considerations in investigating universality classes is also not sharp. For example, Rice (2018) argues that "the goals of the modeler(s) will determine the universality class(es) that are involved in justifying the use of a particular idealized modeling techniques to explain" (2817).

efforts should be directed in assessing the internal validity of the model. Guala (2003) expresses the notion of internal validity as follows: “Internal validity is achieved when the structure and behavior of a laboratory system (its main causal factors, the ways they interact, and the phenomena they bring about) have been properly understood by the experimenter” (1198). By attaining internal validity, scientists avoid establishing dependence relations between structure and outputs within the model that are incorrect. Given that there are intersubjective standards to assess internal validity, a correct explanation of the model’s outputs is a realistically attainable goal.

Third, the explanation of the exemplified outputs in the model is generalized. In this context, generality is a formal quality of an explanation, which amounts to the explanation being cast in categorical terms (cf. Sheredos, 2016: 928). In other words, a general explanation is explicitly conceptualized in terms of ‘types’. The idea is that the generalized version of the explanation does not only explain the exemplified outputs in the model. Rather, this generalized version explains the kind of phenomena of which both the exemplified outputs in the model and the explanandum phenomenon in the target are instances. Another way to put this is to say that the explanation in the context of the model is a token explanation, which addresses token phenomena in the model. What is required is a type explanation, which addresses the pragmatically decided type-explanandum that comprehends both the model’s outputs and the explanandum phenomenon in the target.

The crafting of this generalized version of the explanation requires competence. Scientists must possess a clear notion of the kind of phenomenon that is intended to be explained and make adequate judgments in introducing modifications that extend the scope of the explanation. In practice, this means that some details of the originally crafted explanation may be neglected. Other aspects may be adequately embedded into categorical terms. I refer to this process as “abstraction”. The specific tasks of this abstraction procedure are difficult to prescribe. Rather, which decisions are adequate is a matter that is assessed in context.

Fourth, the generalized version of the explanation is imputed as a model explanation of the explanandum phenomenon in the target. This is justified because the explanandum

phenomenon in the target is a token of the kind of explananda that the generalized model explanation explains. In other words, model explanations are a type-explanation which explain a type-explanandum, from which the explanandum phenomenon-in the target is a token. It is in this sense that the imputed explanation is a potential explanation: it fits the template of an account of explanation with a type-explanandum. Note that, up to this point, the model explanation is imputed to the target without assessment of structural resemblance between model and target. This is because the causal structure of the target is unknown (by construal). In this sense, whether the potential explanation is a correct one for the target is a matter of further research.

A summary of the aforementioned steps is sketched in Figure 6-3. I admit that the steps I just discussed for the operation of model explanations are not sharply distinguished in scientific practice, nor need they follow the order in which I presented them. Instead, they should be regarded as a reconstruction of the explanatory practices of scientists. I proceed to illustrate how this scheme works in the context of my case studies.

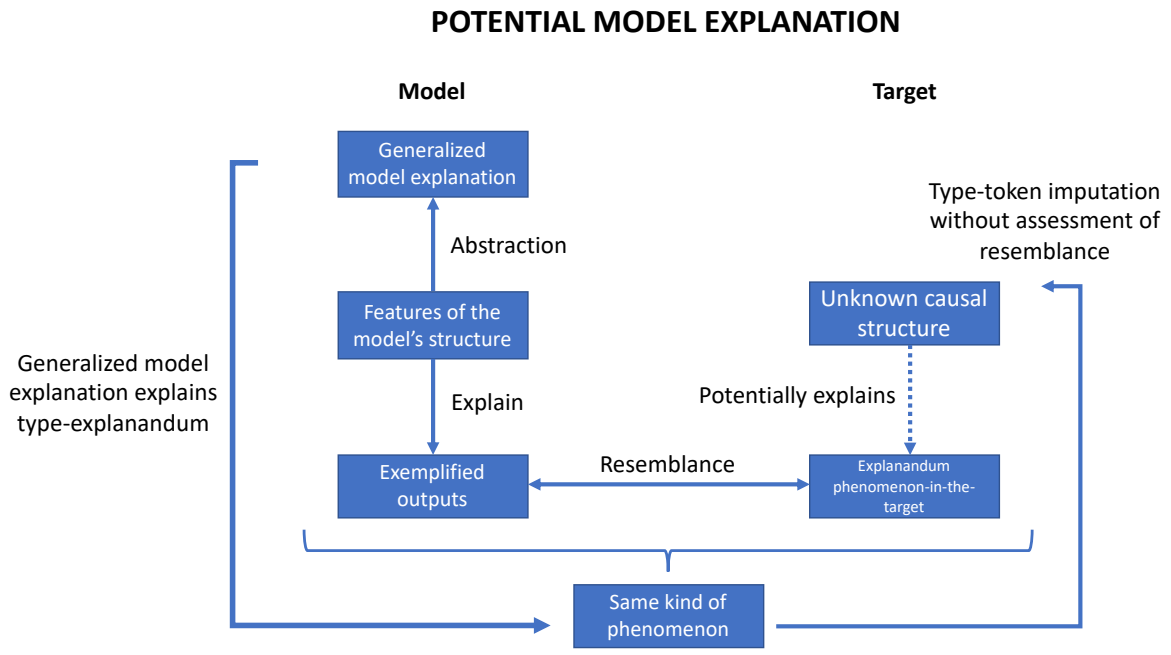


Figure 6-3: Potential explanation with models.

In the context of the OFC model, the explanandum phenomenon in the target is twofold. On the one hand, OFC intend to explain the SOC behavior of earthquakes, which is

epitomized by the Gutenberg-Richter law. On the other hand, they intend to explain a particular aspect of the SOC behavior in earthquakes, namely their non-universality (i.e., different b-values in the power-laws across geographic contexts). The OFC model, as a computer simulation, produces avalanches that follow a robust power-law distribution. Furthermore, the power-law distributions of avalanches in the model instantiate different exponents (i.e., non-universality), depending on the level of conservation in the relaxation stages (captured by α coefficients).

The first task is to establish phenomenal resemblance. Certainly, outputs in a computer program and real earthquakes are quite distinct phenomena. But a relevant, pragmatically decided, resemblance can be found between the exemplified outputs in the model and the target phenomenon. In this case, the relevant resemblance is that avalanches in the model and earthquakes in seismic faults exhibit a power-law distribution of frequency versus size. In addition, the exponents of these power-laws (b-values) are non-universal in both the OFC model and actual earthquakes. Thus, the relevant resemblance between model and target is the robust but non-universal power-law distribution of events' sizes, with 'events' referring to avalanches in the model and earthquakes in the target. This phenomenal resemblance is noted by OFC: "The [OFC] model gives a *good prediction* of the Gutenberg-Richter law […]" (1244; my emphasis). Based on this resemblance, a kind of phenomenon can be established as type-explanandum, namely the robust and non-universal power-law distribution of events' sizes. The Gutenberg-Richter law and the power-law distributions of avalanches in the OFC model are tokens of this type-explanandum.

The second task is to establish an explanation for the robust and non-universal power-law distribution of avalanches in the OFC model. In order to do this, OFC explore dependence relations between the structure of the OFC model and its outputs that are exploitable for explanation. OFC present evidence that suggests a dependence relation between B and α (see Figure 2-16). They state: "The exponent B depends on α" (1246). This dependence does not only account for the variability of B (i.e., its non-universality). It also accounts for the obtaining of robust power-law behavior: power-law behavior obtains for values of α in the range of 0.05-0.25.

It is worth noticing that OFC run several simulations to examine the extent of such dependence. This is also done in their complementary papers (Christensen & Olami, 1992a; Christensen & Olami, 1992b). This is how OFC assess the internal validity of their model. There is strong evidence that suggests that such dependence relation actually exists, and it can thus be used to explain the obtaining of power-laws and the variability in the exponents. In this sense, this explanation has promise to become more than a simply potential explanation and attain a higher epistemic status. I return to this issue below.

To be sure, α coefficients are not the only feature of the OFC model on which the B exponents depend. Just to name another relevant factor, border conditions affect the values of B exponents too (Christensen & Olami, 1992a: 1834). In this sense, a more accurate claim is that the α coefficients *partially* explain the robust and non-universal power-law of avalanches. Or, *ceteris paribus*, α coefficients *actually* explain the robust and non-universal power-law of avalanches. In addition, other structural features are hypothesized to hold dependence relations with the outputs of the OFC model, but fail to have the same explanatory scope of α coefficients. For example, OFC explore the dependence relations between the size of the cellular automaton (L) and its outputs. While the power-law behavior of avalanches is robust for various values of L, the variability of B values does not depend on L (1246).

The dependence relations explored in the OFC model are subordinated to the non-representational interpretation of the OFC vehicle. Because of this, valid but distinct dependence relations can be described between the outputs of the OFC model and its structure. For the most part, the dependence relations in the OFC model, in particular that between B and α, seem to be of a mathematical kind. They are relations between numerical parameters and outputs in the context of an abstract calculus in a cellular automaton. However, as discussed in the previous chapters, OFC's non-representational interpretation of the OFC vehicle is, at times, hybrid: they also explicitly discuss aspects of the OFC vehicle in physico-mechanical terms, borrowed from BK's spring-block model. For example, the α

coefficients are often referred to as “elastic ratios” that express the ratio between the elastic coefficients of springs in the BK model.

The third task is to deliver a generalized version of the explanation crafted in the context of the OFC model. In practice, this means that the dependence relation established in the context of the OFC model – e.g., that between α and B – must be abstracted in such a way that does not only account for the robust and non-universal power-law behavior of avalanches in the OFC simulation. Such abstract version would account for the robust and non-universal power-law behavior of events’ sizes, whether the events in question are avalanches in the OFC simulation or real earthquakes.

I suggest that OFC engage in this kind of generalization by abstracting away from the specifics of their model. The generalized dependence relation is not expressed in terms of numerical parameters in a cellular automaton calculus, nor in terms of elastic coefficients of springs in a spring-block system, nor in terms of the elasticity of rocks. Rather, it is expressed by means of the categorical term “non-conservation.” The notion of non-conservation generalizes the explanatory role of α coefficients, whether interpreted as numerical values in an abstract calculus or as elastic ratios in a spring-block system. As numerical values, α coefficients indicate non-conservation in the sense of establishing a fraction of an a-dimensional numerical value to be added between successive variables in a cellular automaton. As elastic ratios, α coefficients are indicative of non-conservation in the sense of expressing the amount of elastic potential that is redistributed across blocks. The educated guess is that non-conservation is also a categorical term that captures features of the causal structure of the target. However, given that I am testing the case in which no causal knowledge of the target is possessed (i.e., a black box), I leave this discussion for later in the chapter.

The fourth and final task is to impute the generalized dependence relation to the target as a potential model explanation. This imputation is justified by means of a type-token relation between the type explanandum of the general version (i.e., robust and non-universal power-law distribution of events’ sizes) and the token explanandum phenomenon in the

target (i.e., robust and non-universal power-law distribution of earthquakes). As argued above, the imputation of a model explanation as potential explanation does not require an assessment of resemblance between what the general explanation describes and the causal structure of the target.

6.2.2 The Problem: Managing Idealizations in Model Explanations

In the scenario discussed above, scientists have no causal knowledge of the target. As long as this is the case, potential explanations remain explanations in a merely formal sense: they fit the template of an account of explanation and can be imputed to the target based on a type/token relation. In the absence of causal knowledge of the target, the epistemic status of potential model explanations cannot be decided. In other words, given the lack of resources to assess resemblance, it is not clear whether potential explanations provide explanatory knowledge.

However, total lack of causal knowledge about the target is an extreme case. Typically, scientists do possess some amount of causal knowledge about the target, even if elementary. In fact, some amount of causal knowledge is often involved in the design of models that are intended to represent the target. And even in the case in which relevant causal knowledge is lacking, scientists can often implement methods to collect evidence and thus gain some causal knowledge. For example, scientists can design instruments for direct (or indirect) observation, engage in sampling campaigns, and conduct experiments in the actual systems, to name a few options. In other words, target phenomena are not rarely complete black boxes.

In the presence of causal knowledge about the target, the assessment of resemblance between model and target can be extended. That is, the assessment of resemblance is not restricted to phenomenal resemblance, which was presented above as a minimal requirement for potential model explanation. Rather, the assessment of resemblance can now involve judgments of similarity between the causal structure of the target phenomenon and those structural features of the model captured in the model explanation. By engaging in this assessment, the epistemic status of potential model explanations can be boosted.

In principle, the higher the structural resemblance, the higher the epistemic status of the model explanation. In the extreme case of having a perfect replica of the target, studying the replica would be epistemically tantamount to studying the target. Yet, models are not perfect replicas of target phenomena. Rather, scientists conduct research with models which resemble their targets only in some respects, up to a certain degree. This seems to be an inevitable feature of scientific modeling, given the overwhelming complexity of nature and the cognitive limits of scientists as humans (cf. Potochnik, 2017: Chapter 1).

The respects in which models do not resemble their targets are typically referred to as "idealizations." The topic of idealizations has been vastly discussed in the literature, from a plurality of different approaches. In addition, specific matters have been addressed, such as idealizations across scientific disciplines, idealizations in relation to accounts of explanation, and idealizations in distinct kinds of models, to name a few topics. To add to the vastness of the topic, these approaches are commonly assessed across a plethora of case studies, which display the workings of idealizations in actual scientific practices. It is not my intention to survey these various approaches and case studies nor provide an exhaustive, state-of-the-art review of the notion of "idealization". This goes beyond the demands of my argument and – most probably – beyond my capabilities. However, in order to proceed with my argument, I must submit a minimal characterization of "idealization" that fosters a clear discussion. My strategy is to assort the main platitudes across the various approaches found in the literature and adapt them to my overall argument.

Finding these platitudes – however – proves itself a challenging task. Philosophers of science address the notion of "idealization" differently. Dissensus is often found in crucial aspects, such as definitions, classifications, and the epistemic reach of idealizations. As an illustration, consider the widespread disagreement concerning the synonymy (or lack of it) among otherwise related notions such as abstractions, distortions, simplifications, imprecisions, partiality, isolations, omissions, fictions, extractions, approximations, and subtractions, to name a few. Still, I suggest that there is a common thread across the various approaches to idealization: idealizations are aspects of models in which they are *unlike* their targets, with the term 'unlike' differently construed.

Although I do not disambiguate the various forms in which the term 'unlike' is construed, one feature is worth highlighting: 'unlikeness' to the target is pragmatically decided. In practice, this means that idealizations in a model emerge only after a representational interpretation of the corresponding vehicle is in place. And, as discussed in the previous chapter, a representational interpretation of a vehicle requires a previous non-representational interpretation of the vehicle. Given this relation between idealizations and interpretation, it follows that idealizations in a model are also committal.

Thus, as a minimal characterization, I propose the following: idealizations are those features of vehicles that are interpreted – both non-representationally and representationally – in a way that makes them *intendedly* different from their correlates in the target. As a corollary, any inadvertent unlikeness to a user of a model is not part of the idealizations of the model. Note that models that are only used in a non-representational mode cannot be idealized, because there is no specified target with respect to which be unlike.

Characterizing idealizations as intended unlikeness to a target has the unfortunate consequence of making all representational models idealized: models are not interpreted as perfect replicas of their targets. As a consequence, saying that a model is idealized becomes a superfluous statement. Along these lines, Potochnik (2017) submits that idealizations in scientific models are "rampant" and "unchecked." That is, they are widespread and are not – for the most part – intended to be replaced by more accurate descriptions (19).

However, it is not vacuous to claim that models can be more or less idealized, once the criteria for assessment of resemblance are in place. Given that vehicles could be interpreted in ways that involve more or less idealization, it becomes relevant to ask the following question. How much idealization – in what respects and to what degree – still secures a sound surrogative reasoning. As I have argued above, *ceteris paribus*, the more idealized a model is, the less similar it is to the target, and the less sound the resulting surrogative reasoning is.[58]

[58] Ylikoski and Kuorikoski (2010) express similar thoughts in the context of model explanations: "*Ceteris paribus*, a factually more accurate explanation enables a broader range of correct inferences that an

Because of this, scientists often engage in strategies to de-idealize the lessons learnt in the context of a model before imputing them to a target. However, there are three main problems with these strategies. First, they might not be possible to execute. Knuuttila and Morgan (2019) discuss this issue at length. Without going into much detail about their argument, Knuuttila and Morgan question the feasibility of de-idealization tasks, construed as the reversal of idealizations. They focus on two main problems with this approach.

On the one hand, in trying to isolate the idealized features of a model and neutralize them, the very epistemic functionality of the model might be at stake. This is because the epistemic functionality of a model often relies on the integration of its various elements, including the idealizations. In these cases, idealizations are ineliminable: they "appear to be far more 'embedded' or 'intertwined' within the explanation provided by the model" (Rohwer & Rice, 2016: 1138). Along similar lines, Rice (2018) argues that several (if not most) models in science are "holistically distorted representations" of their targets (2811). In this sense, isolations cannot be isolated and de-idealized because the model in question is – itself – an idealization.

On the other hand, even if de-idealizations are possible, they are misleadingly construed as the reversal of idealization procedures. Scientists do not always start with a complex conception of the target phenomenon and construct a model by idealizing such conception. Rather, scientists often start their research by studying simple systems (i.e., models). In attempting to use these systems as surrogates of targets, scientists may need to modify the inferences derived in the context of the model to be applicable to the target. But such modification is not the reversal of idealization. Rather, it is the recontextualization of the inferences derived in the model to be applicable in the context of the target (see "reinterpretation" in Rohwer & Rice, 2016: 1136-7). Knuuttila and Morgan (2019) discuss four types of recontextualization efforts, namely recomposing, reformulating, concretizing, and

explanation incorporating idealizations, presuming that the inferences can actually be drawn without the idealizations in question" (212).

situating (for more details, see ibid: 646-54). Insofar as these efforts are adequately conducted, idealized models may prompt model explanations with a high epistemic status.

The second problem with de-idealization strategies is that, even if they are possible, there might not be a straightforward way to decide these procedures. Given the pragmatic approach to resemblance that I have been advocating, this point may not seem surprising. However, my point is that *even once the criteria for assessing resemblance are decided*, scientists may not be able to univocally decide de-idealization strategies. To be clear, the results of distinct de-idealization strategies may be compared and differently praised in terms of affording model explanations with a higher epistemic status. But I suggest that there is no protocol that could be prescribed in advance that optimizes such epistemic status. I conjecture that scientists may need to resort to sub-optimal procedures (i.e., heuristics), such as various trial-and-error and simply compare the results of various strategies.

As an illustration, consider a case study in economics. Without going into much detail, Svetlova (2013) argues that a particular type of model, known as "valuation" models, are idealized models in the peculiar form of being "under-constrained." A strategy to de-idealize these models is by means of providing a commentary – roughly, a story about how the model fits the world. Even if the standards of resemblance between model and target are fixed, there may be several commentaries that meet such standards. Thus, deciding which commentary is the best de-idealization strategy is not straightforward. Along these lines, Svetlova highlights the role of judgments and expertise in making these decisions (334).

The third problem with de-idealization strategies is that they might not even be desirable: idealizations in models may contribute to the crafting of model explanations with a high epistemic status. In this sense, idealizations – adequately employed – do not preclude sound surrogative reasoning, but actually foster it. This is a point made, for example, by Strevens (2013). He argues that correct explanations (explanations with a high epistemic status) can be – and often are – crafted from idealized models. However, he argues that this is the case insofar as idealizations included in the explanation do not make a difference to the explanandum phenomenon. That is, idealizations in explanations are acceptable as long as they do not do explanatory work. Idealizations play a role in explanations by highlighting the

irrelevance of some factors in explaining a phenomenon. In Strevens's view, idealizations contribute to explanations, but not positively. In fact, if idealizations were removed, the resulting explanation would retain its explanatory power, although with additional non-explanatory information.

Other philosophers argue that idealizations may be positively desirable because they themselves play an explanatory role. Consider for example Bokulich (2011). She argues that an account of model explanations must show how idealizations play a positive explanatory role. Roughly, she argues that genuine model explanations may involve idealizations insofar: i) the counterfactual structure of the model explanations is isomorphic to that of the target phenomenon; and ii) the target falls within the domain of applicability of the model. This allows for idealizations (even fictional entities) to be part of model explanations and play a positive explanatory role.

In light of these arguments, I must qualify my previous claims. Above, I stated: *ceteris paribus*, the more idealized a model is, the less similar it is to the target, and the less sound the resulting surrogative reasoning is. But in actual modeling practices – as the arguments above show – idealized models can be used to craft model explanations with a variable epistemic status, *if the idealizations are adequately managed*. In this sense, the assessment of the epistemic status of model explanations based on idealized models should be conducted in a case by case manner.

To grasp the main aspects of my discussion on idealizations and their relation to epistemic status, consider the BK case. It is clear that there are various idealizations embedded in the BK model. That is, the spring-block system is intendedly interpreted to be unlike seismic faults in various regards. In the previous section, I mentioned that BK highlight two main aspects that call for idealization (or simplification): i) the geometrical configuration of the systems under study (i.e., seismic faults); and ii) the methods of excitation that bring about the earthquakes. Concerning simplifications of the geometrical configuration, the target system is one fault plane which is represented as a discrete set of entities, as opposed to a continuous medium. Concerning simplifications of methods of excitation, a simple

excitation mechanism is in place (the pulling motor), and friction is isolated from other factors for focused study (i.e., different laws of friction can be simulated).[59]

Beyond these two main idealizations, there are several other embedded in the BK model. To name a few: i) focus on one geological fault, as opposed to a system of coupled, interacting faults; ii) a homogeneous geological fault, particularly in terms of its flatness; iii) homogeneity of the driving stress; iv) a one-dimensional array of particles as opposed to a two-dimensional or even three-dimensional, array; v) assumption of Hookean elasticity for the springs; vi) an idealized law of friction, including schematic radiation and viscosity; and vii) assumption of elastic deformation.

Sources from the secondary literature also repair on the idealizations implemented in the BK (and OFC) model. For example, Xia *et al.* (2005) claim that: "Both the CA [i.e., cellular automaton in the OFC model] and nearest-neighbor Burridge-Knopoff models lack several elements that would make them more realistic representations of earthquake systems. In particular, the latter does not include long-range stress transfer, and the long-range CA models do not have inertia and more realistic friction laws" (248501-1). Kagan (2014) also reflect on the shortcomings of the BK model, even though he praises its role as a starting point (16-7). This should suffice to make the case that the BK model is a highly idealized model of real earthquake faults and their behavior.

A similar case can be made for the OFC model. Given that the OFC model is used as a model of earthquakes based on the BK model, one would expect some of the idealizations of the BK model to be retained. A notable example is that seismic faults are also discretized in the OFC model, in the form of a cellular automaton. Beyond the idealizations retained from the BK model, the OFC model resorts to idealizations of its own. A notable example is the design of the interactions among cells. Consider that the BK model retains some physical

[59] These two idealizations are core theoretical principles in the construction of the model. I discuss them as P1 and P2 in previous chapters. Using well-established terminology, these principles can be characterized as follows. P1 (the continuum-discreteness principle) is a Galilean idealization, meant to make the problem tractable. P2 (the isolation of friction) is a minimal idealization, meant to isolate core causal factors that make a difference to the phenomenon. For more details on these types of idealizations, see Weisberg (2013: 99-103).

accuracy by implementing a laboratory model which – as a material model – is subjected to physics. Or by developing a mathematical model based on Newtonian equations of motion. In contrast, the OFC model abstracts away from the details of a physical description. It is based on simple arithmetic operations for the variables. To be sure, these operations retain qualitative aspects of the physical dynamics of a spring-block system, such as threshold-controlled dynamics and the transference of energy to neighboring cells. However, the details of such dynamics are not shaped by theoretical physics nor subjected to physics.

Recalling Knuuttila and Morgan's concerns, some of the idealizations in the BK and the OFC model seem too deep-seated in the functioning of the model to be de-idealized. Consider the case of the discrete structure of both models. It is due to their structure as a set of discrete entities – blocks in the BK model, cells in the OFC model – that these models function as they do. In other words, their discrete structure does not play a merely simplifying role from which the model could prescind. Furthermore, these discrete structures are desirable idealizations. Consider the BK case. BK discuss some preliminary physical intuitions of the dynamics of a continuous elastic string before presenting the spring-block system (see discussion in Chapter 2). However, BK abruptly switch from discussing the continuous elastic string to discussing a discrete version of it (343-4). They do not provide an explicit reason for this switch. However, I conjecture that the reason is that the dynamics of the discrete string is easier to imagine and describe. Actually, after introducing the discrete version of the string, BK proceed to describe its dynamics, as they intuit it unfolds.[60]

Thus, it seems that idealizations embedded in the BK and OFC models are too closely attached to the epistemic functionality of the models, and conveniently so. Thus, if the BK and OFC models are to be used for surrogative reasoning, then an important question to ask is how these idealizations affect the epistemic status of the resulting inferences. While I suggest that this question should be approached in a case by case manner, I would like to

[60] An insightful framework for discussing the idealizations in BK and OFC models is that of the "holistic view" (Rice, 2019; Rice, 2018). The idea is that BK and OFC can be conceptualized as not merely containing idealizations. Rather, the models are themselves idealizations in the sense of *holistically* distorting the target. In this view, models can be used to explain a target because they engage in the same kind of phenomena. In technical terms, the model and the target belong to the same "universality class." Notably, the notion of universality is often used in SOC research to justify the employment of highly idealized models in the study of targets that exhibit SOC behavior (e.g., Pruessner, 2012: 19).

propose a general strategy. In the next section, I address a case in which model explanations can be imputed to the target as genuine explanations, even though the models are highly idealized.

6.3 Attaining Genuine Model Explanations with Idealized Models: a Modal Strategy

6.3.1 Qualified Mappings

In using a model as a surrogate system, a representational interpretation of the model is required. In Chapter 5, I argue that representational interpretation is decomposed in two subtasks, namely establishing the target and establishing the mapping by which elements of the model relate to elements of the target. Still, there seems to be a need for a third task, namely one that qualifies the relations between elements of the model and elements of the target. That is, while the mapping establishes what elements of the model relates to what elements of the target, the qualification establishes how the elements of the model relate to the elements of the target. The idea is that, by having such qualification, scientists gain control over the epistemic status of model explanations imputed to a target. To be sure, the task of qualification does not need to be conceived as the third task of representational interpretation. Instead, it can perfectly be embedded within the second task, namely mapping. My point, however, is to emphasize that mappings do more than just establish a denotative scheme between elements from the model and target. The mappings can and often should include qualifications upon such denotative scheme.

These ideas on qualified mappings are inspired by aspects of the DEKI account of representation (Frigg & Nguyen, 2018; Frigg & Nguyen 2016). Frigg and Nguyen (2018) argue that "properties of a model are rarely, if ever, taken to hold directly in their target systems and so the properties imputed to targets may diverge significantly from the properties exemplified in the model" (217). Frigg and Nguyen aim to build this consideration into their account of representation in the form of "keys." Roughly, keys serve to adequately read a model and translate a property exemplified by a model into a property imputed to the target (as keys on a map). It goes without saying that, in the case of highly idealized models, keys play a central role in employing a model as a representation of a target.

The nature of the keys is left open in the DEKI account. This decision is made based on the myriad of factors that influence them: "the scientific discipline, the context, the aims and purposes for which the model is used, the theoretical backdrop against which the model operates, etc." (ibid: 218). In some cases, the keys cover what has been discussed in the literature as de-idealization tasks (see discussion above). In other cases, the keys impute the properties of the model intact to the target – i.e., "identity" key – which is, admittedly, rarely used. In any case, there are no strict standards for the keys to fulfill. In particular, the keys do not need to comply with resemblance criteria. In fact, the keys are construed in a way that allows them to play a strictly conventional or stipulative role. To be sure, there may be keys that secure a more accurate representation than others (once the criteria for assessment of resemblance is in place), but that is an independent issue. Keys are just an explicit indication of how properties of the model are to be imputed to the target.

In the DEKI account, the keys act upon exemplified properties in a model. I suggest that the notion of keys can be extrapolated to cover other aspects of models intended to be imputed to a target, beyond property ascription. In particular, I consider it fruitful to extend the notion of keys to the imputation of model explanations to a target. Indeed, what is typically imputed to the target as a model explanation is a reworked version of the explanation that explains outputs in a model based on resources of the model.[61] I proceed to discuss a particular kind of key to be applied to model explanations, especially those based on highly idealized models: modal qualifications.

6.3.2 Preliminaries on Modalities of Explanations

For the purposes of my dissertation, modalities can be roughly characterized as modes according to which the veracity of a proposition is assessed (see "alethic" modalities in Kment, 2017). The idea is that propositions can be (approximately) true or false depending on how

[61] Note that my scheme of potential explanation above implicitly resorts to the notions of keys. As I argue above, in order to impute a model explanation to a target as a potential explanation, the explanation of the model's outputs must be abstracted. This abstraction is conducted in such a way that the resulting generalized explanation explains the type-explanandum of which the outputs in the model and the target phenomenon are tokens.

they are qualified. For example, consider the qualification terms 'necessarily', 'possibly', impossibly', and 'contingently' and the proposition 'I am a geologist'. While it is true that I am a geologist, it is contingently true and not necessarily so. In addition, it could have been possible for me not to be a geologist. Furthermore, these qualifications may have a logical, metaphysical, or physical scope of application. For example, while it its physically impossible to move faster than the speed of light (in accordance with the laws of physics), it is not logically impossible (i.e., it is not a logical contradiction for an object to move faster than the speed of light).

I provide a basic scheme of relations between modalities (Figure 6-4). This scheme is used as the basis upon which I elaborate more detailed views. First, I submit a partition of possibilities and impossibilities, i.e., what could be the case and what cannot be the case. As a partition, these two categories exhaust all the space of modalities and are mutually exclusive. Second, another partition is introduced, namely that between actualities and non-actualities, i.e., what is the case and what is not. In assembling these two partitions – (im)possibilities and (non-)actualities – I submit the following relations. All actualities are possible. Non-actualities may be possible or impossible. Third, a finer grained partition is introduced within the actualities and the non-actualities. This partition distinguishes the contingent from the necessary. All contingencies are possible. Necessities can be possible (and thus actual) or impossible (and thus non-actual).

<table>
<tr><td colspan="3">Possible</td><td>Impossible</td></tr>
<tr><td colspan="2">Actual</td><td>Non - Actual</td><td></td></tr>
<tr><td>Contingent</td><td>Necessary</td><td>Contingent</td><td>Necessary</td></tr>
</table>

Figure 6-4: Modalities and their relations.

To be sure, these distinctions may not hold in other – less naturalistic – contexts. For example, one could make the case that not all that is impossible is necessarily so: Imagine a physical event which is not allowed by the laws of physics (i.e., impossible). And now consider that the governing laws of physics are contingent (not necessary) to the particular evolution of the universe. This consideration pushes me to restrict the way in which I talk about these modalities. For the purposes of my dissertation, I constrain my discussion to physical

modalities, as opposed to more general metaphysical or logical modalities. In case I need to talk about logical or metaphysical modalities, I express that explicitly.

Still, physical modalities may be an ambiguous notion. Consider physical possibilities. They may refer to what is physically possible given the physical constitution of the world. Or they may refer to what is physically possible given our contingent knowledge of physics. This is a Pandora's box that I try to keep relatively closed. For the purposes of this dissertation, I assume that scientific knowledge is the best arbiter to decide physical modalities. In this sense, a physical modal claim in science is not simply a claim about what scientists think is the case. It is a claim about how the world is (but see my discussion on "plausibility" below).

As a last preliminary remark, I discuss the notion of plausibility. Note that a plausibility partition is not depicted in Figure 6-4. The reason is that I intend to construe plausibility as a distinct form of modality, orthogonal to those depicted in Figure 6-4. This distinction has been discussed in the literature in terms of physical and epistemological modalities. The modalities depicted in Figure 6-4 are meant to be physical. In contrast, plausibility is a modality that qualifies the veracity of a proposition in light of the available evidence to a particular agent. The way I intend to construe "plausibilities" allows them to be physically contingent or necessary, actual or non-actual, and possible or impossible. This depends on the reasoning capabilities and evidential knowledge available to the agent assessing the modality. This is what I mean by orthogonality: epistemic and physical modalities vary independently. To be sure, this is a controversial construal: the relations between epistemic and physical modalities are a matter of active research today (e.g., see Fischer & Leon, 2017).

To illustrate the (putative) orthogonality between plausibility and the physical modalities, consider two extreme cases: the plausibility of impossible phenomena and the implausibility of actual phenomena. On the one hand, impossible phenomena might seem plausible to a perfectly rational epistemic agent, insofar the rational epistemic agent is, to some extent, ignorant of the governing physical architecture of the universe. I would argue that, for instance, mythological explanations of natural phenomena refer to impossible phenomena that are deemed plausible by the epistemic agents of the time and place. On the other hand, actual phenomena may be deemed implausible. Sure, for the most part, actual

phenomena are commonly deemed plausible, given that there is commonly evidence that supports them. However, the contingent epistemic state of an agent might make actual phenomena seem implausible. In particular, this is the case whenever the agent's background knowledge conflicts with the actual or the available evidence is biased towards states of affairs that refute the actual.

The reason to elaborate on modalities is that they allow for a refined characterization of scientific explanations. More explicitly, the content of explananda and explanantia in scientific explanations can refer to what is, could, must, or could not have been the case. This is to say that explananda and explanantia are modal notions. Explananda and explanantia can refer to things that are actual – either contingently or necessarily – or non-actual – while still possible or impossible. The modalities of explananda and explanantia in science can span any of these options. In the following, I shape my discussion in terms of modalities of scientific explanation *simpliciter*, although it should be noted that it applies in particular to "model" explanation, as a species of scientific explanation.

6.3.3 Modal Qualifications of Explanations

I structure my discussion around three modal qualifications of explanations, namely "how-possibly," "how-actually," and "how-plausibly" explanations. These modal notions are not mutually exclusive nor do they exhaust all the varieties of modalities. The main reason why I focus on these three notions is that they have become a standard scheme for debates on the modalities of explanation. This is particularly the case in the new mechanist literature, at least since the seminal paper by Machamer *et al.* (2000: 21).[62]

Discussing the epistemic status of how-possibly, how-actually, and how-plausibly explanations is a contentious matter: it is not clear what is meant by how-possibly, how-plausibly, and how-actually explanations. Contrasting accounts of these modal notions have been delivered in the literature, beyond the boundaries of the new mechanist literature. In

[62] "Scientists in the field often recognize whether there are known types of entities and activities that can *possibly* accomplish the hypothesized changes and whether there is empirical evidence that a possible schemata is *plausible*" (MDC, 2000: 17; my emphasis).

light of these conceptual problems, a preliminary task to be performed is the presentation of my own stance. I carve my position in relation to established positions in the literature.

a) Actual Explananda

I restrict my discussion on the modal qualifications of explanations to explanations with actual explananda. As a consequence, the terms 'how-possibly explanation', 'how-plausibly explanation', and 'how-actually explanation' are intended to be a statement concerning the modal status of the explanantia. This restriction is motivated by the nature of my case studies. Both BK and OFC study and try to explain actual phenomena, namely the robust and non-universal power law distribution of naturally occurring earthquakes. Beyond my case studies, explaining actual phenomena is, for the most part, the common explanatory enterprise of science. Scientists mostly aim to explain what the case is, as opposed to what the case is not.

It is worth mentioning some exceptions to this general claim. On the one hand, scientists may be interested in explaining phenomena that they conjecture to be actual, which turn out to be, in fact, non-actual. This is a common scenario, for example, in early exploratory stages of inquiry. Even more so, the search for an explanation of explananda mistakenly thought to be actual often leads to the realization that the explananda is, in fact, non-actual. On the other hand, scientists may be interested in explaining non-actual phenomena as a way of understanding the actual. Weisberg (2007) best describes the merits of this enterprise:

> *"It is also possible to study a model of a phenomenon that is known not to exist. A. S. Eddington once wrote: "We need scarcely add that the contemplation in natural science of a wider domain that the actual leads to a far better understanding of the actual" […] Agreeing with Eddington, R. A. Fisher explained that the only way to understand why there are always two sexes involved in sexual reproduction is to construct a model of a three-sexed sexually reproducing population of organisms […] Constructing a model of such a phenomenon is the only way to study it because, by stipulation, the phenomenon does not exist. Modelers are often interested in phenomena such as three-sex biology, perpetual motion machines, or non-aromatic cyclohexatriene because, insofar as we can understand*

why these phenomena do not exist, we will have gained a better of [sic] understanding of phenomena that do exist" (223).

b) Relations Between 'Possibility' and 'Actuality'

In accordance with Figure 6-4, possible explanantia can be either actual or non-actual, while impossible explanantia are only non-actual. In this sense, actual explanantia are a proper subset of possible explanantia. How-possibly explanations are explanations whose explanantia are possible. How-actually explanations are explanations whose explanantia are actual. As a corollary, all how-actually explanations are how-possibly explanations. However, not all how-possibly explanations are how-actually explanations. That is, how-possibly explanations may have explanantia that are possible but non-actual. Although it is the case that all how-actually explanations are how-possibly explanations, I suggest that an informative account of modal qualification should make the actuality of possible explanantia salient. In other words, explanations with actual explanantia should be referred to as how-actually explanation, even though the qualification "how-possibly" also applies to them, strictly speaking.

This position is not universally shared. In particular, it conflicts with Dray's (1957) account of how-possibly explanation. In his account, an explanation is a how-possibly explanation if the questioner assumes the impossibility of the explanandum and, by means of a possible explanans, she learns that she is mistaken about such assumption. The problem I have with this account is that an explanation might qualify as a how-possibly explanation, even though it might refer to the actual events that caused the explanandum. In my view, if a possible explanans is also actual, then the resulting explanation is a how-actually explanation: the actuality of the explanans explanation should be made salient.

In a widely quoted example, Dray illustrates his notion of a how-possibly explanation. The example refers to a radio transmission of a baseball game. The commentator narrates the following play: "It's a long fly ball to center field, and it's going to hit high up on the fence. The center fielder's back he's under it, he's caught it, and the batter is out" (Dray 1957, 158). For the common listener, this seems like an impossible event, because the fence was over six

meters high. A how-possibly explanation – in Dray's sense – shows how it was possible for the center fielder to catch the ball at this height. In Dray's example, a how-possibly explanation of this event is that the catcher climbs a ladder and reaches a platform from which he catches the ball. This explanation shows that it is not impossible to catch the ball six meters high. However, this also happened to be the actual way in which the catcher caught the ball. Such actuality, I suggest, should be made salient. This is why I would refer to this as a how-actually explanation.

The problem with Dray's account is that his notion of impossibility is not precise enough. If an explainer regards an explanandum as an impossible phenomenon *simpliciter*, then there is no point in engaging in explanation. The qualification is that an explainer might consider an explanandum phenomenon contingently or physically impossible. Depending on which sort of impossibility is considered, the explanatory task surveys the possibility of the explanandum in an alternative category. In the example above, the explainer regards the explanandum as contingently impossible. That is, given the way in which baseball is commonly played, there is no possible way to make that catch. However, it is not physically impossible. The physical possibility of the explanandum motivates the search for an explanans. In this particular case, a minor change in the contingent state of affairs – the addition of a ladder – suffices to update the assumed contingent impossibility into an actual possibility.

c) How-Plausibly Explanations

How-possibly and how-actually explanations are modal forms of explanation that are meant to make a physical claim about the corresponding explanantia. In contrast, how-plausibly explanations are modal forms of explanation that are intended to make an epistemic claim. How-plausibly explanations are explanations that do not seem to conflict with known facts about the explanandum phenomenon, given the available evidence to a certain agent. Furthermore, an explanation can be more or less plausible, depending on the amount of evidential support. That is, how-plausibly explanations admit degrees.

Qualifying an explanation as "how-plausibly" is contingent to the state of knowledge and evidential resources. But this does not mean that the qualification of an explanation as

"how-plausibly" is utterly subjective. Instead, I suggest that the relevant background knowledge and evidential support are those accepted by a community of scientists working within a common research program. In this sense, "how-plausibly" is an intersubjective modal qualification.

Empirical support may increase the plausibility of an explanation, but it does not modify its physical modal status as actual or non-actual. In this sense, I disagree with accounts that pose how-possibly and how-actually explanations as end members of a continuum (e.g., Brandon, 1990; Resnik, 1991). According to these views, the difference between how-possibly and how-actually explanations is a quantitative distinction of empirical support, which comes in degrees. The idea is that most how-possibly explanations tend to involve speculative elements in their content. The role of empirical support is to dispel these speculations, turning how-possibly explanations into how-actually ones. In this case, how-possibly explanations are simply unbeknownst how-actually explanations.

These views portray the modalities of how-possibly and how-actually as if they were epistemic modalities. I do not adopt this construal. The gathering of evidence does not change the physical status of the speculative elements in a how-possibly explanation. Certainly, the speculative elements in a how-possibly explanation can be felicitous or infelicitous. In this sense, whenever the speculative elements are felicitous, collected evidence may eventually support the actuality of how-possibly explanations. But, whenever the speculative elements are not so felicitous, no amount of evidence will be able to turn a how-possibly explanation into a how-actually explanations. In particular, this is the case for non-actual how-possibly explanations, which I proceed to introduce.

d) Non-Actual How-Possibly Explanations

Non-actual how-possibly explanations are explanations that resort to possible explanantia that are not the case (cf. "merely possible" in Vaidya, 2015). Non-actual how-possibly explanations pose a problem to Brandon and Resnik's views (see above). Non-actual how-possibly explanations are how-possibly explanations that are not in a continuum with how-actually explanations. In other words, non-actual how-possibly explanations cannot

become how-actually explanations, regardless of the amount of evidence collected. It is important to note that non-actual how-possibly explanations are not restricted to explanations with speculative elements. Instead, scientists can assume knowingly non-actual posits. This is not an uncommon scientific practice: studying the non-actual is an enlightening path to come to learn about the actual.

My notion of non-actual how-possibly explanations conflicts with other accounts of how-possibly explanation available in the literature. For example, consider Salmon's views on how-possibly explanation (1989): "[A] how-possibly question does not require an actual explanation; any potential explanation not ruled out by known facts is a suitable answer" (137). The problem with Salmon's account is that potential explanations, as an answer to a how-possibly question, must be in line with known facts. This leaves out a number of scientific explanations with possible but non-actual explanantia that conflict with known facts. I presume that what Salmon describes is more suitably referred to as "how-plausibly explanation," i.e., how-possibly explanations that do not conflict with the evidence. In this sense, it seems that Salmon agrees with Brandon and Resnik in their epistemic approach to how-possibly explanations. I have already made the case that this approach has significant limitations.

e) How-Actually Explanations

There are physical actualities: natural phenomena actually occur. And there is an actual set of causal and dependence relations that make such physical actualities happen. Still, it may go beyond scientists' capabilities to deliver a full account of all the actual relations involved in bringing about actual phenomena, in all their complexity and details. Such full account may also go beyond scientists' purposes. In this sense, it remains critical to decide realistic standards to assign the qualification "how-actually" to an explanation. I proceed to discuss these standards in terms of completeness and accuracy criteria (cf. criteria for "correct explanation" in Strevens, 2013).

I submit that how-actually explanations are not required to describe completely and fully accurately the actual bringing about of explananda phenomena in all their complexity

and details. On the one hand, the qualification of "how-actually" is attributed to an explanation that at least describes those aspects that are explanatorily relevant to the obtaining of the explanandum, i.e., the difference makers. Any additional non-explanatory information is not relevant for assigning the qualificative of "how-actually". This is the "completeness" criterion. On the other hand, the difference makers should be described accurately enough so they actually account for the obtaining of the explanandum. This is the "accuracy" criterion. Note that the accuracy criterion does not apply to non-difference makers. In this sense, a how-actually explanation can idealize aspects that are not relevant to the explanandum or neglect these aspects altogether.

In practice, the completeness and accuracy criteria could be relaxed. The notion of "how-roughly" explanation captures this attitude (cf. Glennan, 2017: 71). The idea is that a how-roughly explanation is informative enough about the actual causes of the explanandum, without fully satisfying the completeness and accuracy criteria. In this sense, how-roughly explanations are just how-actually explanations with a more pragmatic attitude. Indeed, the completeness and accuracy criteria seem to be excessively demanding. In practice, scientists take some explanations to be the actual explanations, even though they may not live up to the completeness and accuracy criteria. In other words, some how-actually explanations are, strictly speaking, only partial and approximate explanations.

This pragmatic attitude is not intended to make "how-actually" an epistemic modality, i.e., relative to the capabilities of the explainer. What makes an explanation a 'how-actually' one is decided by how the world is. The point is that, in attempting to convey what is actual, an explanation may fall short of describing this objective part of the world in all its complexity and details. In other words, 'how-actually' is a physical modality but, given human limitations in putting together explanations, this physical modality is ascribed to an explanation based on pragmatic judgments.

f) Levels of Abstraction and Modal Qualifications

I suggest that how-possibly, how-actually, and how-plausibly explanations can be crafted at different levels of abstraction, i.e., they can include more or less explanatory details

or be crafted in more or less categorical terms. As a consequence, the degree of abstraction does not decide the modal status of an explanation. In this sense, I agree with Bokulich (2014) who argues that providing an explanation with more or less details does not straightforwardly decide its status as how-possibly or how-actually explanations. She claims: "What is fundamental to the how-possibly/how-actually distinction *at any level of abstraction* on my account, however, is precisely what one would expect: that is, whether or not the mechanism represented in the model is *in fact* the mechanism responsible for the production of that phenomenon in nature" (Bokulich, 2014: 335, my emphasis).

An important case that provides support to this position is presented by Forber (2010). He argues that how-possibly explanations may be crafted at different levels of abstraction. On the one hand, there are "global" how-possibly explanations, which are crafted at a high level of abstraction. This is the case because they are directed to general, often idealized, target systems. Because of this, global how-possibly explanations are shaped by formal constraints, often posed by mathematical theories and analytical techniques. On the other hand, there are "local" how-possibly explanations which are directed to specific target systems. Because of this, local how-possibly explanations are supposed to be shaped by empirical constraints derived from the study of specific systems.

It is worth noticing that global and local how-possibly explanations are distinct in their relation to evidence. Local how-possibly explanations (along with how-actually and how-plausibly explanations) embody an inductive pattern. That is, their status as "local how-possibly," "how-plausibly," and "how-actually" is inductively assigned as confirmatory evidence is being accumulated. In contrast, global how-possibly explanations are deductively derived: once the theoretical framework is set and formal rules are established, possibilities can be deductively judged. In this sense, global how-possibly explanations embody a sort of *a priori* knowledge. I embrace this conclusion and argue that how-possibly explanations, directed to a general target, need not respond to evidence of specific systems.

Reydon (2012) suggests that global how-possibly explanations are not full-blown explanations in the sense of having no particular phenomena as explananda. Or, at least, that is what he suggests before hedging his claim later in his paper: "Even if I'm right in claiming

that not all of Forber's kinds of explanations explain particular phenomena [this includes global how-possibly explanations], they still might be explanations aimed at different kinds of explananda" (309). I consider this latter claim an important clarification. I propose that, instead of thinking of global how-possibly explanations as explanations with no explananda, they can be thought of as explanations with a general, highly abstract, explananda.

For other philosophers, the level of abstraction influences the modal status of explanations. For example, Persson (2012) argues that there is a conception of how-possibly explanations in which they reveal the mechanism responsible for bringing about the explanandum phenomenon described in the explanandum, but only partially. In other words, relevant (i.e., explanatory) details of the inner workings of the posed mechanism are not explicitly stated. Persson's approach to how-possibly explanation seems to collapse into what Machamer *et al.* (2000) refer to as mechanisms sketches or schemata. A mechanism schema is an abstract (i.e., not detailed) description of a type of mechanism, which can be filled in with more detailed descriptions of token activities and entities. A mechanism sketch is an abstract description that has gaps and even might lack bottom-up entities and activities. In this sense, a mechanism sketch is qualitatively more abstract than a schema: while a schema lacks token-information, a sketch lacks type-information. In other words, a mechanism sketch is an incomplete mechanism schema.

6.3.4 Deciding on an Adequate Modal Qualification for Genuine Model Explanations

Now that the tool of modal qualifications is available, I suggest that highly idealized models can be used for genuine model explanations, without the need of engaging in de-idealization tasks. A model explanation based on a highly idealized model can be imputed to a target explanandum as a genuine explanation, provided that an adequate modal qualification is attached to it. There are two aspects of this claim that I would like to examine more closely. First, I discuss in which sense the so imputed model explanations are genuinely explanatory of their targets. Second, I sketch a few strategies to decide adequate modal qualifications.

First, I tackle the genuineness of model explanations based on highly idealized models. Typically, model explanations are considered genuinely explanatory if they are true, or approximately true, of their target explananda. But what is true should not be confounded with what is actual. Besides actual facts about the target, there are modal facts concerning the target (Vaidya, 2015). In this sense, genuine model explanations do not need to have actually true explanantia. Model explanations may have plausibly true or even possibly true explanantia. If the right modal qualification is attached to these model explanations, they can be imputed to the target as genuine model explanations, which state true propositions about the plausibilities and possibilities of the target.

It seems that, for a considerable number of philosophers, genuinely explaining is just explaining actually. The bias towards actuality can be exposed in three accounts of genuine model explanations with idealized models, surveyed in depth by Bokulich (2011). In the first account, model explanations are genuinely explanatory if they correctly reproduce the actual mechanisms operating in the target (Craver, 2006).[63] In other words, a genuine model explanation must describe the mechanisms responsible for the explanandum phenomenon completely and accurately. This implies that idealized models cannot be used for genuine model explanations. In the second account, model explanations based on idealized models may be genuinely explanatory but only if the idealizations are harmless. That is, idealizations must not obstruct whatever is doing the actual explaining (Elgin & Sober, 2002; see also Strevens, 2013). In the third account, model explanations based on idealized models may be genuinely explanatory but only if the idealizations playing an explanatory role can be de-idealized to recover the actual explanation (McMullin, 1985). Similar views can be found elsewhere in the literature on the genuinely explanatory role of model explanations based on idealized models (e.g., Alexandrova & Northcott, 2013).

I oppose these views. I suggest that model explanations based on highly idealized models may be genuinely explanatory in the sense of providing knowledge about the actual phenomenon. But this knowledge does not need to be about the actual ways in which the phenomenon occurred. Instead, it may include knowledge about the plausible or even

[63] Craver (2006) makes this point in the context of mechanistic model explanations.

possible ways to bring about the actual phenomenon. Thus, how-possibly and how-plausibly model explanations may be genuinely explanatory about the actual target, either by providing explanatory knowledge concerning the target's space of possibilities or its state of evidential warrants, respectively.

In this sense, modal qualifications are not merely a sort of "hedging." They do not merely express hesitation. Rather, modal qualifications are affirmative: they express a definite claim about a modal fact. In other words, by prefixing a model explanation with these qualifications, the model explanation affirms something true about the target explanandum. A how-possibly model explanation states something true about the possibility space of the target explanandum. And a how-plausibly model explanation states something true about what we are warranted to believe given the available evidence. Note that I argue that how-possibly and how-plausibly model explanations *may be* genuinely explanatory about the target. This depends on whether the modal qualification is adequate for the model explanation in question.

It is worth noticing that my approach to genuine model explanations is broader than Bokulich's approach. In a later paper, Bokulich (2012) considers two fictional models which do not explain how-actually a target phenomenon occurs: i) the classical periodic orbit model explaining conductance properties of quantum dots; and ii) the epicycles model explaining the retrograde motion of Mars. Without going into much detail, Bokulich argues that the former fictional model explains genuinely, while the latter does not. The reason for this is that the former, according to Bokulich, is a fictional model that "adequately" represents the relevant features of the target. Allegedly, the latter fictional model does not do this. But, as she says, what counts as an adequate fictional representation is negotiated by the relevant scientific community (734).

This is where Bokulich and I differ. I consider the epicycle explanation of the retrograde motion of Mars genuinely explanatory *if qualified as a how-possibly explanation*. Even if scientists do not find this an adequate fictional representation today, it does not stop it being a less possible explanation than it was in Ptolemy's days. There is an important lesson here: while adequateness in representation may be negotiated across a scientific community,

possibilities are not so negotiable. Following Forber (2010), possibilities can be deduced from a theoretical framework. In this sense, within a same theoretical framework, standards for adequate representation can be negotiated while physical possibilities are obliged by the framework.

The second topic I want to examine more closely is the strategies to decide adequate modal qualifications for model explanations. I suggest that these strategies can be characterized as the pitching of the highest epistemic status of a potential model explanation (see section 6.2.1). This pitching is based on a series of pragmatic judgments concerning similarity between model and target. As I argued above, a basic pragmatic judgment of resemblance between model and target should already be in place, namely phenomenal resemblance between model and target. Otherwise, a model explanation cannot be considered a potential explanation of the target explanandum in the first place: there would be no minimal tether between model and target. This phenomenal resemblance translated into the establishment of a type-explanandum.

After this, the next step is to secure a how-possibly status to the potential model explanation. I suggest that this is tenable if the internal validity of the explanation of the model's outputs based on the model's structure is established. If this is the case, the potential model explanation is a genuine how-actually explanation of the model's outputs. Furthermore, if the abstraction of this explanation is conducted competently, then its abstract version is a how-possibly explanation of the type-explanandum (see Figure 6-3). This is the case by definition: by describing the actual bringing about of the model's outputs, the model explanation describes a possible bringing about of the kind of phenomena of which both the model's outputs and the target explanandum phenomenon are an instance. To put it bluntly, a how-actually explanation of a token is a how-possibly explanation of the corresponding type.

After this, the challenge is to claim that the model explanation is not only a how-possibly explanation of the type-explanandum, but also a how-possibly explanation of the token explanandum phenomenon in the target. This is the type/token imputation I discuss above (6.2.1). The logic behind imputing a modal qualification from type to token is best

captured by Roca-Royes (2017). She delivers an account of *de re* possibilities which are imputed to concrete objects based on similarity judgments. The basics of her argument are captured by what she calls a "naïve logic" (226):

i) We know that *x* actually F.
ii) And, we know that *x* and *y* are similar.
iii) Then, we know that *y* possibly F.

The crux of the matter is that *x* informs us about *y* because they are similar. As I have argued above, this similarity between *x* and *y* enable us to construe them as tokens of a same kind of phenomenon. Because of this, *x* and *y* can be taken as epistemic counterparts. In my case, the "naïve logic" is the following:

i) We know that P_M actually [is explained by E_M].
ii) And we know that P_M and P_T are relevantly similar [i.e., they are tokens of a same kind].
iii) Then, we know that P_T possibly [is explained by E_M].

where P_M: Model's outputs; P_T: explanandum phenomenon in the target; E_M: model explanation. As a consequence, the potential model explanation can be imputed to the target explanandum phenomenon as a genuine how-possibly model explanation.[64]

An expected criticism towards my extrapolation of Roca-Royes's thesis is that the epistemic counterparts in my naïve logic are not concrete objects, but phenomena. I think this is not a problem for the obtaining of the naïve logic described above. If anything, Roca-Royes calls for attempts to extrapolate her cases to more complex ones: "[A] lot more remains

[64] Rice *et al.* (2018) express similar ideas, describing them as a form of inference to the best explanation. They argue that: "the best explanation for why the model and the real-world system behave in the same way [i.e., are relevantly similar] is that the pattern of interest is stable across systems with different physical components and interactions. In other words, scientists use the fact that the consequences of the idealized model are approximated by the behaviors of the real-world system(s) as evidence that the model and real-world system(s) are members of a class of systems that will display similar patterns of large-scale behavior despite (perhaps drastic) differences in their physical details" (14).

to be done than has been accomplished here. The hope of the project is that we can nonetheless generalize beyond the simple cases to cover more complex cases in saliently analogous ways" (235). Still, if the criticism remains, a modified version, which uses concrete objects, may be a solution:

i) We know that M actually F.
ii) And we know that M and T are similar.
iii) Then, we know that T possibly F.

where M: the BK or OFC model; T: seismic faults; F: [instantiates the content of E_M]. In this case, M and T are concrete objects. And the similarity between M and T can be assessed in line with the mapping provided in the representational interpretation. While it is the case that M and T are foreseeably dissimilar in various respects (all models are idealized), what is important is that M and T *do* hold relevant similarities.

In order to attribute a higher epistemic status, the "black box" of the causal structure of the target must be open. That is, some causal knowledge of the target must be in place. This way, assessments of causal resemblance may be conducted, and the epistemic status of the model explanation can be established. Elements of the model explanation are compared to their correlates in the target, according to the mapping decided in the representational interpretation of the vehicle.

At least one thing is foreseeable about the assessment of resemblance: the match is less than perfect. Models involve idealizations that make then unlike their targets. Because of this, potential model explanations can hardly be imputed as genuine how-actually explanations of the target phenomenon. However, in practice, scientists may take model explanations to be genuine how-actually model explanation if they satisfy intersubjective standards of completeness and accuracy (see discussion above). And even if the "how-actually" qualification seems too demanding, it can be relaxed into a "how-roughly" qualification. This way, partial and/or approximate model explanations may be deemed a high epistemic status.

A potential model explanation can also be deemed a genuine how-plausibly explanation if the available evidence and background knowledge prompt the qualification. As I argue above, plausibility is an epistemic status which depends on the background knowledge and evidential support accepted by the relevant community. In this sense, this modal qualification is adequately attributed to a model explanation only relative to a specific context. In my case studies, the relevant specific context takes the form of a research program.

I proceed to explore how modal qualifications can be attributed in the context of model explanations with the OFC model. Recalling my discussion at the end of Section 6.2.1, the type-explanandum in question is the robust and non-universal power-law distribution of size events. The potential model explanation is that the variability of α coefficients cause the non-universal but robust power-law behavior of avalanches in the OFC model. Given that this explanation has been tested in several simulations, with varying assumptions, its internal validity is established. Hence, this explanation is a how-actually explanation of the outputs of the model. Based on what I argue above, this how-actually explanation of the model's outputs is also a genuine how-possibly model explanation of the type-explanandum (a how-actually explanation of a token is a how-possibly explanation of the corresponding type). In its abstract version, however, the model explanation does not make reference to α coefficients, which are idiosyncratic to the model. Instead, the abstract version of the model explanation makes reference to the general notion of non-conservation.

According to my discussion of Roca-Royes's naïve logic, the abstract version of the model explanation can be imputed to the token target phenomenon as a genuine how-possibly explanation. This is the case because the target phenomenon and the model's outputs are tokens of a same kind of phenomenon. That is, they share relevant similarities. Most notably, the relevant similarities are captured by Jensen's SDIDT conditions for SOC phenomena (1998), or Watkins *et al.*'s genotype of SOC phenomena (2016) (see Chapter 2). Take Jensen's SDIDT account, for example. The BK model, the OFC model, and seismic faults are similar in having a **S**lowly **D**riven, **I**nteraction **D**ominated, **T**hreshold controlled dynamics. These are genuine issues of resemblance that prompt the employment of the OFC model as epistemic counterpart. Accordingly, the naïve logic prompts the type/token imputation. Thus,

OFC have a model explanation of the robust and non-universal power-law behavior of earthquakes based on non-conservation. This model explanation is, at least, a genuine how-possibly model explanation.

In order to pitch a higher epistemic status, OFC must have some causal knowledge of seismic faults and the occurrence of earthquakes. This is indeed the case: OFC are expert scientists who have background knowledge and evidence about earthquake occurrence. This way, assessments of resemblance can be conducted between the content of the model explanation and the causal structure of the target in accordance with the accepted mapping. The match is less than perfect, as expected. More explicitly, the different degrees of non-conservation in the context of the OFC model are instantiated by the α coefficients. In the case of the target, non-conservation is instantiated by viscoelastic or plastic components of deformation of rocks plus seismic radiation. Here is the crux of the matter: α coefficients in the model and radiation in the target are unlike each other in a significant regard. On the one hand, the α coefficients are the ratios of elastic constants established between *discrete* interacting blocks. On the other hand, non-conservation in the form of viscoelastic or plastic deformation and/or seismic radiation occurs in the *continuum* of a seismic fault and rocks. In a few words, the key explanatory element – namely non-conservation – is instantiated discretely in the model and continuously in the target.

This unlikeness between model and target suggests that a "how-actually" qualification for the model explanation is out of the question. And the "how-plausibly" qualification also seems inadequate given the refuting evidence and conflict with the background knowledge. However, it is important to recall that resemblance is assessed pragmatically. Because of this, it is worth noticing how OFC perceive the explanatory power of their model. For example, Christensen and Olami (1992a) claim that the results of simulations with different α coefficients provide "some explanation" of the observed variability of B exponents (1835). In addition, they claim that there are "good reasons to believe that this simplistic picture [the OFC model] has a *real connection* to the *actual* fault dynamics that leads to earthquakes" (ibid: my emphasis). In this sense, it seems reasonable to argue that OFC conceive of their

explanation as a genuine how-roughly model explanation (i.e., a how-actually explanation which fails to fulfil strict criteria of accuracy and completeness).

Certainly, this appraisal may be debated by other scientists with other standards for the pragmatic assessment of resemblance. But even if the explanation is qualified as a non-actual how-possibly model explanation, it still affords genuine explanatory insight into the target. In a way, what OFC do is surveying the space of possibilities associated with a type-explanandum (robust and non-universal power-laws). Thus, OFC contribute to the enlargement of a class of explanations that are worthy of further research. What I am trying to convey here is a notion of a how-possibly explanation very much in line with Forber (2010)'s notion of global how-possibly explanation. These how-possibly explanations are shaped by formal considerations, typically embodied in mathematical terms. They embody a kind of mathematical necessity that makes them significant epistemic achievements by themselves, even though their relation to particular targets may be unclear or not even intended to exist.[65]

SOC computer models, and the derived model explanations, are – I suggest – this kind of explanatory achievement: How-possibly explanations in a formal sense. As Pruessner (2012) argues, several SOC computer models "are not intended to model much except themselves […] in the hope that they display a certain aspect of SOC in a particularly clear way" (3). It is true that the OFC model is introduced as a model of earthquakes. But it is also true that, as a mathematical formalism, it is studied in itself, regardless of what it may represent. While the OFC model can be used to deliver how-possibly explanations of the behavior of earthquakes, it can be used to explore how-actually explanations of its own behavior, based on its mathematical structure.

6.4 Précis

[65] Cuffaro's notion of "algorithmic" explanation (2015) is also relevant. He develops this account in the context of explanations in computer science. As its name suggests, this account of how-possibly explanations explains by describing the pathways available for an algorithm to follow or, in his own words, "by explicitly specifying the set of possibilities open to them [i.e., the algorithms]" (746). OFC's explanation is a how-possibly model explanation in the algorithmic sense: it is based on the algorithm of the cellular automaton and it specifies the possibilities open to the simulation.

The main theoretical theses of this chapter are:

- Thesis 1: Scientific models can be used as surrogate systems to study targets of interest because they are accessible and knowable. In particular, they can be used to deliver model explanations.
- Thesis 2: Phenomenal resemblance between model and target is a minimal condition for potential model explanation. This resemblance prompts the adoption of a type-explanandum of which the model's outputs and the target explanandum phenomenon are tokens.
- Thesis 3: Highly idealized models can be used for genuine model explanation, provided that an adequate modal qualification is in place, namely how-possibly, how-plausibly, or how-actually (in practice, how-roughly). These qualifications are decided based on pragmatic assessments of resemblance between model and target.

The main lessons learnt from my case studies are:

- The BK and OFC models are surrogative systems used to craft model explanations of the occurrence of earthquakes. In particular, the OFC model is explicitly used to explain a target phenomenon, namely the robust and non-universal power-law distribution of earthquakes in terms of their size and frequency.
- OFC provide a model explanation which – in its abstract version – explains the robust and non-universal power-law distribution of events' sizes in terms of variability in non-conservation.
- This explanation is a genuine how-possibly model explanation of the target explanandum phenomenon. However, it is a non-actual one, given that model and target are relevantly dissimilar in their instantiations of non-conservation. In the OFC model, non-conservation is defined discretely. In the target, non-conservation occurs continuously.
- Nonetheless, OFC seem to appraise this explanation as a genuine how-roughly model explanation, based on their pragmatic assessment of resemblance.

***

7 Scientific Understanding with Models

In this chapter, I provide a *theoretical synthesis* of the topics discussed in the last three chapters, viz., explanatory commitments, interpretation in models, and model explanations. This synthesis is crafted in the context of a firmly established – although at times messy – theoretical framework in the philosophy of science, namely that of "scientific understanding." This framework is a particularly expedient one to craft this synthesis due to the central role of understanding in the scientific enterprise. This is captured in the following quote by Bridgman (1950): "The act of understanding is at the heart of all scientific activity; without it any ostensibly scientific activity is as sterile as that of a high school student substituting numbers into a formula" (72; cited in de Regt, 2017: 1). By using this framework, the impact of my previous discussions becomes more conspicuous. Given that the task I engage in is a theoretical synthesis, I do not explore my case studies much in this chapter. Rather, my intention is to rephrase some of the insights acquired in the last chapters to shed a different light upon the case studies. In the next chapter, with this theoretical synthesis at hand, I resume my exploration of the case studies.

In practice, this synthesis amounts to crafting an *account of scientific understanding* that capitalizes on the notions discussed above. I expect this account to be a significant contribution to the ongoing debates in the literature in at least three senses. First, it connects the notions of "explanatory commitments" and "model interpretation" intimately to scientific understanding. While models and explanations have been widely discussed in the literature on scientific understanding, they rarely have been addressed in conjunction with the notions of commitments and interpretations. Second, the resulting account serves as a response to influential accounts of scientific understanding available in the literature. Furthermore, I suggest that my account outperforms alternative accounts in specific regards to be discussed below. Third, the account is motivated by – and tested in – novel case studies, namely the BK and OFC models (see Chapter 8).

There are various challenges involved in crafting an account of scientific understanding. In particular, there are two kinds of problems that can be referred to as the "plurality" problem and the "fuzziness" problem. The plurality problem has two manifestations: one in the context of science, the other in the context of philosophy. On the one hand, there is widespread but dissimilar and often vague employment of the term 'understanding' in the sciences. Accordingly, it is difficult to provide an account of scientific understanding that captures the various ways in which 'understanding' is conceived by scientists. On the other hand, there is a variety of different – often conflicting – philosophical accounts of scientific understanding. Thus, any new proposal is susceptible of being met with skepticism or even destructive criticism.[66]

The plurality problem may be mitigated by assuming a pluralist attitude towards scientific understanding. In this sense, the existing variety of accounts of scientific understanding may be regarded as a fertile landscape that provides various resources for the discussion. Even more so, the very variety of accounts of scientific understanding may reflect the multifaceted nature of scientific understanding. Or it may well reflect the existence of different types of scientific understanding.

However, assuming a pluralist attitude is not always straightforward. This brings me to the fuzziness problem. The theoretical framework of scientific understanding is still work in progress. That is, the various uses of the term 'understanding', the various types of understanding and their interrelations have not yet been settled (cf. Baumberger *et al.* 2017: 5). A particular problem in this regard is the existence of different typologies of understanding

[66] The existing variety of accounts of scientific understanding is a rather recent circumstance. For decades, scientific understanding was neglected as a topic of serious discussion in the philosophy of science. This is noted by Henk de Regt, one of the champions of contemporary philosophical discussion on scientific understanding. He claims that, although several scientists and philosophers talk about understanding, the lack of a proper account of understanding inspired him to write his 2017 book on the matter (de Regt, 2017: ix-x). The lack of a proper account of understanding can be – at least partly – explained by two forms of disregard among philosophers of science. On the one hand, an important scholarly tradition used to attribute the pursuit of understanding to the humanities (*verstehen*), while the natural sciences were in the business of achieving explanations (*erklären*) (see e.g., Droysen, 1868). On the other hand, as understanding became recognized as an aim of the natural sciences, it was mostly construed as a derivative one, namely a psychological by-product of explanation. Only in the last two decades, scientific understanding has become a stable and rather popular research program in the philosophy of science and epistemology. For more details on this historical reconstruction, see Baumberger *et al.* (2017).

whose types are vaguely characterized and overlap with each other. This does not only make it difficult to navigate the theoretical framework. It also makes pluralism an aspiring stance given that the plurality of approaches to scientific understanding is not clearly schematized. Thus, given the plurality and fuzziness problems, the scope of this chapter needs to be spelled out and some stipulations are required.

My account of scientific understanding acknowledges two types of understanding, commonly referred to as "explanatory" and "objectual" understanding (but cf. Khalifa, 2013). These types are consistently discussed in the literature, although with varying construals. As a minimal characterization, Baumberger *et al.* (2017) propose the following. Explanatory understanding is understanding why something is the case.[67] Objectual understanding is understanding some subject matter or domain of things (5). Admittedly, these characterizations are rather broad. But this is exactly what is intended: they are meant to capture general consensus concerning these types. Naturally, this leaves various aspects open for discussion.

Given the dissent on how to deal with these types of understanding, I resort to the following strategy. I begin by presenting my accounts of explanatory understanding (EU) and objectual understanding (OU). These accounts are based on my appraisal of the various accounts discussed in the literature and the lessons drawn from my case studies. They are not meant to provide a definition of scientific understanding but rather criteria to attribute it. It is worth noticing that, according to this construal, EU is a special case of OU:

(EU) An agent has explanatory understanding of a target phenomenon with a model if and only if the agent possesses the ability – and its object – of crafting and/or using a genuine model explanation of the phenomenon, regardless of its modal qualification.

[67] As a clarification, explanatory understanding arguably goes beyond merely understanding *why*, including also understanding *how* a phenomenon occurs. This means that explanations that afford understanding of a phenomenon do not merely *indicate* the difference makers for the obtaining of the phenomenon, but also *describe them in more detail*. In this sense, the difference between explaining why a phenomenon occurs and explaining how it occurs can be characterized as a difference in grain of description. As an illustration, assume a new mechanist approach to explanation. Explaining why a phenomenon occurs amounts to indicating the responsible mechanism. Explaining how the phenomenon occurs involves a fine-grained description of such mechanism, including its entities, activities, and organization, along with an account of how all of them together are responsible for bringing about the phenomenon.

(OU) An agent has objectual understanding of a target phenomenon with a model if and only if the agent possesses the ability – and/or its object – of crafting and/or using an effective representational model of the target.

The rest of this chapter can be read as an elaboration of these accounts. To begin with, I proceed to deliver preliminary remarks that apply to both EU and OU. These remarks amount to a delineation of the scope of these accounts and the presentation of a structural feature common to both, namely the "ability–object of ability" dichotomy. Then, I discuss EU and OU in terms of the components of the aforementioned dichotomy, i.e., the object of abilities and the abilities themselves. This discussion prompts me to contrast my views with other influential accounts of explanatory and objectual understanding available in the literature. By engaging in this comparison, I am able to locate my stance within the debate while highlighting its contributions.

7.1 Preliminary Remarks: Structure and Scope

7.1.1 Structure: the "Ability–Object" Dichotomy

To motivate the following discussion, consider a seemingly frivolous piece of syntactic analysis: 'understanding' is a gerund. Roughly, a gerund is a verb form that functions as a noun. In this sense, gerunds embody a dichotomy between verb and noun, between actions and things. As a gerund, 'understanding' conveys a dichotomy between actions and the objects of such actions. That is, understanding may be conceived as an activity and/or as the object that such activity leads to or makes use of.

It is worth delivering a first approach to the actions and objects that figure in understanding, even though I discuss these issues in further detail below. The "activity" component of scientific understanding is typically discussed in terms of "abilities". Roughly, abilities are *competent actions*. This articulation has an important implication: not just any action affords understanding, but only the right kind of action. Thus, the term 'ability' introduces a normative element into understanding. In line with the scope of my argument,

this normative element is decided at the level of research programs. Abilities can be conceptualized as "upstream" or "downstream" abilities. Upstream abilities are those that lead to the obtaining of an object. Downstream abilities are those that are enabled by means of possessing an object.

The objects of abilities that figure in scientific understanding can be broadly construed as a sort of *epistemic content*. This content may be what results from exerting an ability successfully, or it may be what enables an ability to be exerted successfully. (Note that I am not using the term 'object' to denote what is being understood.) The object of abilities – i.e., the content – is distinct for EU and OU. In EU, the object of abilities is model explanations. In OU, the object of abilities is representational models. There is a further distinction between the objects of abilities in EU and OU. In EU, the object of abilities is construed as a *genuine* piece of knowledge. In OU, the object of abilities is construed as *effective* content. I discuss this distinction in detail below.

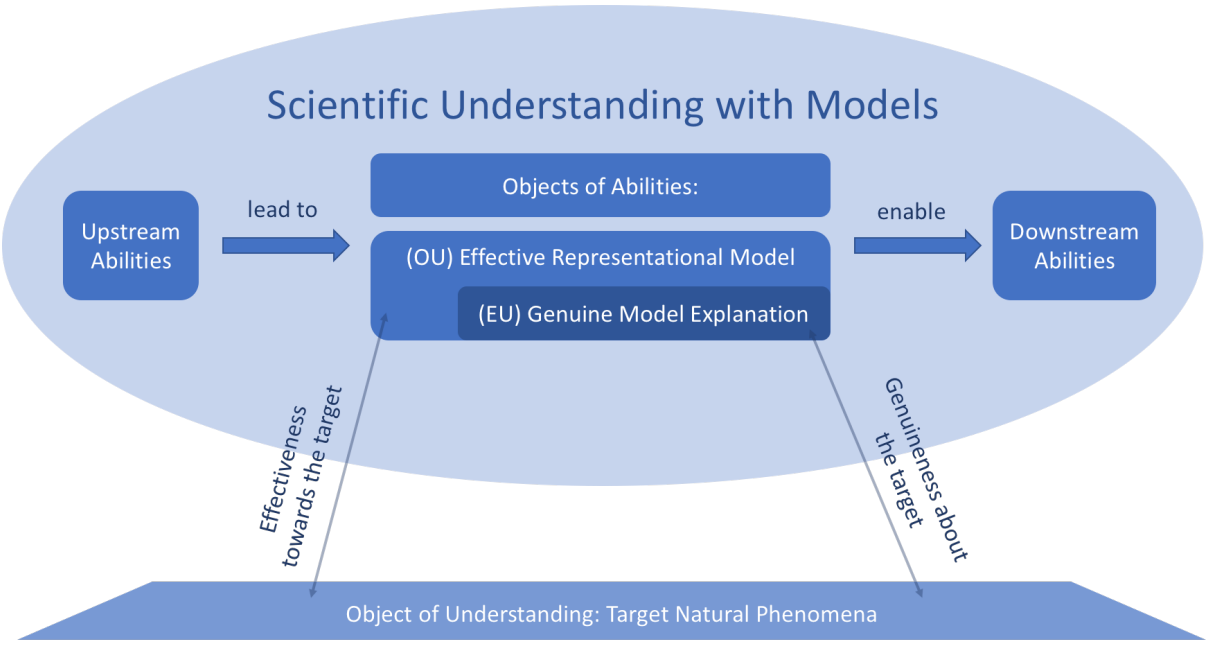


Figure 7-1: Ability–object dichotomous structure of scientific understanding with models.

A great part of the debates on scientific understanding consists in characterizing the abilities and objects that figure in understanding. While there is great dissent on the specifics, most accounts of scientific understanding manifest the ability–object dichotomy. To support

this claim, I will briefly present three accounts of scientific understanding that embody the dichotomy, although with dissimilar approaches. These accounts are due to de Regt (2017), Lawler (2019), and Wilkenfeld (2017).

First, consider de Regt's account of explanatory understanding (2017). He argues that explanatory understanding is "both a means to, and an end of, explanation" (16). More explicitly, the end is the attainment of understanding of phenomena, which amounts to "having an adequate explanation of the phenomenon" (23). The means is the understanding of theories, which amounts to "being able to use [theories]" to craft explanations (23). Note that the understanding of theories (i.e., the means) is non-explanatory: it does not rely on having an explanation. Rather, the understanding of theories is equated with the ability to use theories to craft explanations. The ability to use theories is an upstream ability which leads to explanations. In this sense, de Regt's account of explanatory understanding embodies an "ability–object" dichotomy. More explicitly, explanatory understanding involves an ability – namely the ability of using theories – and its object – namely, the resulting explanation.

Second, Lawler (2019) distinguishes between the *process of obtaining understanding* and the *obtained understanding* (17). She makes this distinction to elucidate the role of falsehoods in objectual understanding. Roughly, the idea is that falsehoods may enable the obtaining of objectual understanding. But falsehoods are not themselves part of the content of such understanding. Rather, falsehoods are used to extract factual information about a subject matter. This is why she refers to her account as the "extraction view." Lawler introduces and discusses a variety of "extraction methods" to obtain true information from falsehoods (20). As she diagnoses, these methods are not arbitrary: they are legitimized by being successful in the extraction of true information about a subject matter (20). In this sense, the methods of extraction are abilities. In exerting these abilities, the content of understanding is obtained. This content is the object of abilities. Lawler argues that both aspects – the process and the product – figure in scientific understanding, even though only the later figures in the content of understanding. In this sense, she also attributes a dichotomous nature to understanding.

Third, Wilkenfeld (2017) suggests that understanding can be assessed along various independent dimensions ("Multiple Understanding Dimensions" or MUD thesis). However, he focuses his discussion on two dimensions, namely "intelligibility" and "representational accuracy." These two dimensions, Wilkenfeld argues, cover most discussions concerning the evaluation of understanding. On the one hand, a state of understanding is more or less intelligible if the possessor of the understanding is more or less able to engage in certain tasks as a result of having such understanding (i.e., "downstream" abilities). On the other hand, the representational accuracy metric is concerned with the representational content of understanding and its relation to a target. This content can be more or less accurate as a representation of the target. The representational content is the object of abilities, i.e., it is the resource that can be implemented into skilled action. In this sense, these two dimensions of assessment embody the projected ability–object dichotomy.

I construe abilities and their objects as constitutive of scientific understanding with models. But this claim comes with an important caveat. In EU, both the ability and its object are *necessary* for understanding. In OU, abilities and their objects are components of understanding in an inclusive disjunctive relation. That is, at least one of these two components – i.e., abilities *and/or* objects – must be present to attribute understanding. In other words, each component is *sufficient* to attribute OU. In some cases, we attribute objectual understanding to agents who possess both ability and object. In other cases, objectual understanding may be readily attributed to an agent based only on the presence of ability or only on the presence of the object. To be sure, the latter agents may be attributed less objectual understanding than the former ones. But in no case is objectual understanding attributed to an agent who lacks both the abilities and their objects. The reasons for distinguishing the necessity or sufficiency of abilities and their objects in EU and OU will become clearer later in the text. However, a first approach to these reasons can be stated here. I construe EU as a special case of OU. EU is stricter than OU in the sense of having a particular kind of content as object of ability (model explanations) and requiring both abilities and their objects.

Construing abilities and their objects as components of scientific understanding (even in the inclusive disjunctive sense of OU) is a contentious decision. Other accounts in the

literature focus exclusively on one or the other component, or even none of them. As an illustration, consider Wilkenfeld (2017), which I briefly discussed above. According to his MUD thesis, intelligibility and representational accuracy are not constitutive of understanding but rather dimensions of assessment of the quality of understanding. This means that they may be part, or not, of a state of understanding, in a variable degree. I agree with Wilkenfeld in adopting intelligibility and representational accuracy as metrics of the quality of a state of understanding. However, I consider that these dimensions – more broadly construed as abilities and their objects – must be present in some degree. It is not clear to me what would be the target of MUD assessment without the presence of abilities and/or their objects.

Now that the ability-object dichotomy has been established, I proceed to present some constraints to the scope of my account of understanding based on EU and OU.

7.1.2 Scope: Three Delimitations

There are three considerations that restrict the scope of my accounts of explanatory and objectual understanding. First, both EU and OU are limited to the understanding that is gained by means of *modeling*. This restriction is a methodological decision: I intend to capitalize on my previous discussions on the topic of modeling. In this sense, I remain open to the existence of other sources of understanding, even though I am skeptical about this prospect. In any case, I will not pursue this issue.

Second, I restrict the object of understanding (i.e., what is being understood) to *target natural phenomena*. This is a key reason why my account focuses on explanatory and objectual understanding: they are types of scientific understanding whose object is (aspects of) the world, whether construed as explananda or subject matter. In the context of my discussion, natural phenomena should be conceived broadly as the various kinds of mind-independent entities susceptible to being studied scientifically. In this sense, phenomena include – but are not restricted to – events, concrete objects, processes, states of affairs, properties, and so forth. Some mind-dependent entities, such as concepts and theories, are often presented as objects of understanding (e.g., “do you understand what ‘mitosis’ means?” or “Do you understand Einstein’s general relativity theory?”). I admit that the term

'understanding' can be meaningfully applied to concepts, theories, models and other mind-dependent entities. (In fact, I discuss the "understanding of theories" and "understanding of models" as forms of ability below.) However, my point is that the understanding of mind-dependent entities is not, in itself, EU nor OU.

Third, my account of scientific understanding does not cover the phenomenological dimension of understanding. This dimension amounts to the psychological experiences brought about by understanding, often referred to as the "feeling of understanding" or "a-ha experience." This does not mean that I deny the existence of these psychological experiences as part of understanding. Rather, the avoidance of these experiences in my account responds to a methodological decision: I intend to deliver an *externalist* account of understanding. This means that my account is meant to provide criteria that can be used to attribute understanding to another agents. Possession of abilities – as a criterion to attribute understanding – does part of the work. Ylikoski and Kuorikoski (2010), alluding to Wittgenstein, express this idea clearly: "understanding should not be conceived as a special mental state, but a publicly attributed behavioral concept akin to an ability" (205).[68] Thus, one way to attribute understanding to agents is by assessing the success of their abilities, which are meant to lead to content or are enabled by it. In addition, possession of the object of ability – as a criterion to attribute understanding – must also be publicly accessible. This means that possession of the content of understanding is not a mere subjective personal feeling. It is a form of explicit knowledge. Such explicit knowledge is expressible. This allows other agents to judge and eventually share this knowledge across an epistemic community (see discussion below).

Having established the dichotomic nature of understanding and delineated the scope of my accounts, the next step is refining what is meant by ability and object in the context of

[68] A similar consideration prompts me to avoid the notion of "grasping" as constitutive of understanding. Grasping is often posed as a relation between mind and world (e.g., Strevens, 2013). This approach makes the obtaining of understanding not assessable to an external agent. This is why I prefer to avoid including the notion of grasping into EU or OU. Still, some philosophers explicate 'grasping' in externalist terms, i.e., in terms of abilities. For example, Grimm (2017) explicates 'grasping' as the ability to make modal inferences, i.e., the ability to foresee changes in one part of a system due to changes in other parts (217).

EU and OU. In the following, I proceed to elaborate my views on this topic while contrasting them with alternative approaches in the literature.

7.2 The Object of Abilities in EU and OU

In this section, I proceed to elaborate on the object of abilities in explanatory and objectual understanding. I submit that the distinction between EU and OU is, precisely, a difference in the object of abilities that figure into understanding. In EU, the object of abilities is *genuine model explanations* (regardless of their modal qualification). In OU, the object of abilities is *effective representational models*. Note that I have discussed my views on genuine model explanations and representational models above (see Chapters 5 and 6). In this section, my intention is to construe these objects as components of EU and OU respectively, a necessary one in EU, a sufficient one in OU. Before discussing what distinguishes one object from the other, I elaborate on their commonalities and relations.

7.2.1 Objects of Ability in EU and OU as Content plus Metacontent

To begin with, representational models and model explanations figure into objectual and explanatory understanding, respectively, as a form of propositional content. On the one hand, the propositional content of a model is the sets of propositions that are true about it or in accordance to it, i.e., propositions that are true in the "world of the model" (cf. Morrison & Morgan, 1999: Ch. 2). On the other hand, the propositional content of a model explanation is the set of propositions upon which the model explanation is built. As a corollary, the propositional content of a model explanation is a proper part of the propositional content of the corresponding model. To be clear, the idea is not that models are just sets of propositions (but cf. Rohwer and Rice, 2016: 1130). Rather, the assumption is that, even if models are not propositional models, there is a set of propositions that are true of them.

Given that models are interpreted vehicles (see Chapter 5), their propositional content depends on the interpretation that they embody. A natural concern may be that the content of models ends up being arbitrary. This is not the case. The fact that a model's content depends on its embodied interpretation does not make it arbitrary because there are, at least,

two main kinds of constraints in interpreting. First, there are objective constraints posed by the mind-independent nature of the vehicle. Second, there are intersubjective constraints posed by the relevant community of scientists (e.g., those belonging to a research program). These two kinds of constraints prevent "ad-hoc" or "wishful" interpretations.

This propositional content *enables* representational models and model explanations to figure in scientific understanding as objects of abilities. However, in order for the models and model explanations to actually figure in understanding, an epistemic agent must relate to their content. Propositional content is, in principle, mere information or "stuff" detached from an agent (cf. "knowledge-as-stuff" in Collins, 2010: 6). I suggest that an agent relates to this content in terms of "possession" of the object of ability. I use the term 'possession' in a technical sense. Possession, as a concept, has three components, which I proceed to introduce.

Consider the case in which understanding is attributed to an agent based on their possession of objects of abilities, whether model explanation in EU or representational model in OU. Firstly, this means that the agent who understands – i.e., the understander – *knows* (part of) the propositional content of the model explanation or representational model in question. In other words, content figures into understanding by means of the understander knowing it. However, this knowledge is not tantamount to understanding. If that were the case, there would be no difference between propositional knowledge of a model and understanding of the target.

A second kind of knowledge is involved in understanding, i.e., involved in the possession of the object of ability. This is a form of metaknowledge, i.e., knowledge about knowledge. More precisely, understanding is not mere knowledge about propositional content, but also knowledge about how pieces of propositional content hang together.[69] Along these lines, Kvanvig (2009) argues that understanding involves a "grasp of the structural relationships (e.g., logical, probabilistic, and explanatory relationships) between the central items of information regarding which the question of understanding arises" (97). I refer to the

[69] These pieces may be singular propositions or sets of them.

relations between pieces of content collectively as "metacontent." Thus, the object of ability in EU and OU is the propositional content of a model explanation or representational model, respectively, plus their metacontent (see Figure 7-2). Understanders possess the object of ability insofar as they know the content and metacontent of a model explanation (in EU) or representational model (in OU).

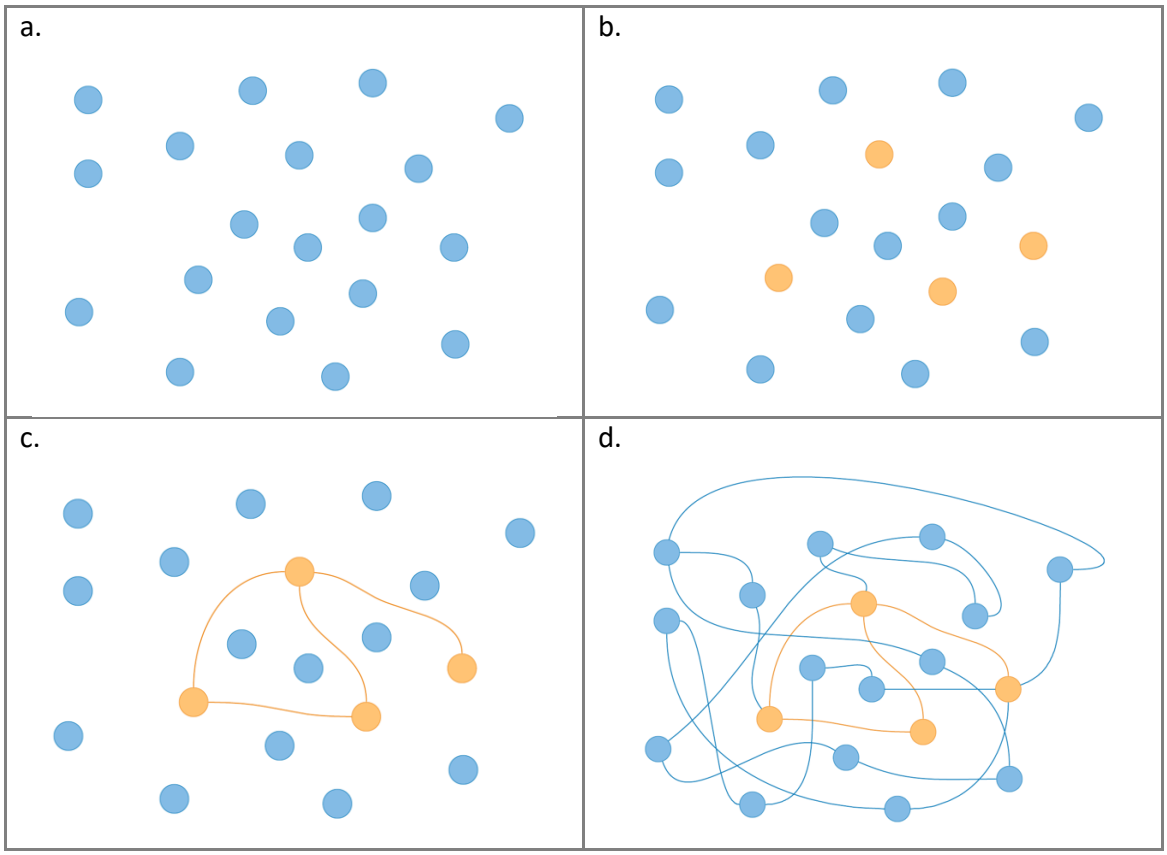


Figure 7-2: Representations of: a) A model's propositional content. Each blue node represents a piece of propositional content – i.e., a single proposition or a set of them – that is true about the model; b) A model explanation's propositional content. Orange nodes are singled out as those pieces of propositional content used for model explanation. Note that the propositional content of model explanations is a proper part of the propositional content of representational models; c) Object of ability in EU. It amounts to the propositional content used in a model explanation plus its metacontent (i.e., the relations that make the propositional content hang together in the model explanation); d) Object of ability in OU. It amounts to the propositional content of a representational model plus its metacontent (i.e., the relations that make the propositional content hang together in the model). Note that the object of ability in OU may include the object of ability of EU. This makes OU broader than EU.

There is still a problem left. Agents do not – and, arguably, cannot – relate to the whole body of propositional content of a model. I elucidate this point. Consider a proposition p that is true about a model. Then, p is part of the propositional content of the model. But then so it is the proposition [p v q] for any proposition q. Thus, the propositional content of a model

is – in principle – infinite and remains, for the most part, unknown. An agent can only relate to part of the propositional content with a particular metacontent at a time. But this "active" content is not readily available to an agent. Rather, it is the agent who must assemble the active content. And it is the agent who must establish – and keep adjusting – its metacontent.

This is a problem that deserves some elaboration. My strategy is to argue that these tasks – assembling the active content and managing its metacontent – are abilities that are endogenous to the object of abilities. More emphatically, these abilities are not to be conceptualized as upstream nor downstream abilities because they are not in the business of creating content nor capitalizing on it. These abilities consist in knowing how to navigate the content of an already existing model. The successful conducting of these abilities leads to the configuration of the object of abilities, i.e., explicitly known content plus its metacontent.

There are two major constrains in executing this navigation. First, the agent must be able to recognize and extract those portions of the propositional content that concern the target phenomenon (the object of understanding). This active part of the propositional content of a model is the knowledge to which agents are committing with regards to the target. In the case of OU, the active content might relate to the target phenomenon quite broadly. However, in the case of EU, the content must be skillfully assembled in a way that affords explanatory information about the target explanandum. As Jebeile and Graham Kennedy (2015) submit: "In order for a model to be successfully explanatory, the scientists who employ the model must be *able* to draw explanations from it" (386; my emphasis). The particular form of this information (causes, probabilities, counterfactuals, etc.) depends on the explanatory commitments of the agent assembling the content. But recall that a model embodies an interpretation that aligns with a scientist's explanatory commitments. Thus, a competent scientist should be able to assort explanatory information that concerns the target phenomenon and aligns with her commitments.

Second, given that the active content is a form of commitment, the agent must be able to put it together coherently (see discussion in Chapter 3). To secure – and further – coherence in the body of active content, agents require more than just choosing the right content. They also need to establish the right relations among portions of it. In this sense,

agents require a form of modal knowledge concerning the content and metacontent of a model. More explicitly, the agent must possess knowledge about the counterfactual structure of a model and about the impact of alternative configurations on the overall coherence of a body of propositional content. On some occasion, this kind of modal knowledge may be a priori. That is, scientists may know how to assemble coherent configurations, just by reflecting on the content and their potential relations. However, on other occasions, the necessary modal knowledge is only acquired by exploring the model, intervening it, and learning from this process.

In sum, the object of abilities – whether as model explanation in EU or as representational model in OU – can be characterized as follows. The object of abilities is a body of propositional content – i.e., the active content – plus the relations that hold pieces of this content within the body together – i.e., its metacontent. Understanding is attributed to an agent, based on possession of the object of abilities. Possession of the object of ability amounts to the following: i) the agent knows how to assemble the active content and adjust its metacontent in a way that affords a coherent body (i.e., modal knowledge); ii) the agent explicitly knows the active content; iii) the agent explicitly knows the relations holding between pieces of content within the body, i.e., the metacontent. Having addressed what is common between the object of abilities in EU and OU, now I discuss what is different between them.

7.2.2 The Effectiveness Condition on the Object of Abilities in OU

The main difference between the object of ability in EU and OU is the epistemic relation that the corresponding propositional content holds to the target of understanding. The propositional content of a model explanation in EU needs to be modally true about the target, i.e., the model explanation must be genuine. In contrast, the propositional content of a representational model in OU needs not be true, at all, about the target. Instead, the requirement is that the propositional content of a representational model in OU leads to successful applications, i.e., it must be effective. I proceed to clarify these claims. I begin with OU.

The propositional content of a representational model is the set of propositions that are true about the model or in accordance with it. But this propositional content may not be true outside the world of the model. For example, consider the proposition 'geological faults are discretized surfaces'. This proposition is true in accordance with the OFC model. But it is not true about actual geological faults. Accordingly, believing such proposition leads to knowledge about the world of the model but not knowledge about the actual target. My purpose is to spell out an attitude towards the propositional content of models that affords a meaningful relation towards the model's target.

I suggest that the way in which propositional content figures in understanding is the following. As stated above, understanders possess the object of ability insofar as they know an active content, its metacontent, and have modal knowledge concerning the model. But in the case of OU, an additional attitude towards this content must be in place which turns it into a meaningful attitude towards the target. My suggestion is that understanders must know the active content of a model and *accept it* as propositional content of the target. In other words, knowledge about a model becomes objectual understanding of the target by means of acceptance. The notion of "acceptance" contrasts with that of "belief". While belief in a proposition p is a disposition to represent p as true, acceptance of p means to take p as a premise without conviction of its being true. For the purposes of my argument, it is important to highlight the relation between acceptance and action. Acceptance of a proposition p is a disposition to act in accordance with p, to take p as a basis for action (e.g., in the case of a working hypothesis). In this sense, acceptance is active: it involves agency, willingness and ability. In other words, one cannot accept that what one does not intend to use nor does not know how to act upon. For further details on "acceptance", see Elgin (2017: Ch. 2).[70]

The distinction between acceptance and belief is important given that portions of the propositional content of representational models – while true in the world of the model –

[70] The distinction between belief and acceptance is not always straightforward. To illustrate this point, consider the ideal gas law. In using the ideal gas law, one accepts that particles in an ideal gas do not interact. However, such acceptance could be expressed as believing that particle interactions do not make a difference to the ideal gas law. This example shows that there are cases in which an acceptance can be transformed into a belief, making the distinction unclear. This is problematic because, while it is rational to accept falsehoods that have instrumental value, it is irrational to believe falsehoods.

may be false as content about the target. Accordingly, believing the propositional content of a model leads to knowledge about the model, but not necessarily to knowledge about the target. Acceptance fares better in this regard: accepting propositional content, even if false about the target, puts the understander in a position that is *susceptible* to achieving meaningful relations towards the target.

I highlight the term 'susceptible' because acceptance of the content of a model, by itself, does not necessarily lead to understanding of the target. In addition, the content must satisfy a condition that leads to meaningful relations to the target. What I have in mind is that the content of a model must be able to lead to practical applications or epistemic achievements. This is called the "effectiveness" condition. A model is said to be "effective" if and only if its content *may be* used in a reliable successful way (cf. de Regt & Gijsbers, 2017: 55).[71] Effective models are the object of abilities in objectual understanding.

There are four aspects of the effectiveness criterion that deserve further commentary. First, I emphasize the words 'may be' in the previous definition. The idea is that a model's effectiveness is a dispositional feature. A model is effective in the sense of being the right kind of tool for a certain purpose. But such purpose needs not be carried out. In other words, downstream abilities need not materialize for a model to be effective. Second, the effectiveness of a model is not intrinsic to it, but rather relative to the intended purposes and the skills of the agent. An effective model provides content upon which a competent agent in a particular context could capitalize. In other words, effectivity is a highly contextual notion. Third, I characterize success as the attainment of aims. I suggest that some criteria must be in place to decide whether the aims in a given circumstance are met. This makes success a kind of judgment. Within the scope of my argument, it suffices to consider some intersubjective criteria decided by a relevant community. In the context of my argument, the relevant community amounts to those scientists working in a common research program. Note that the criteria for success needs not be "all-or-nothing." Specific ventures can be more or less successful in a gradual manner. For the purposes of my argument, these ventures

[71] Along these lines, Potochnik (2017) submits that understanding requires "*successful* mastery, in some sense, of the target of understanding" (94; my emphasis).

amount (for the most part) to practical applications and epistemic achievements (see "downstream abilities" below). Fourth, the request for reliable success is intended to eliminate accidental or lucky success. In other words, a model is not effective if it leads to success by mere chance. An effective model, in the right hands and in appropriate situations, leads to success *ceteris paribus*.

To move on to the issue of "genuineness", I must clarify my views on the relations between effectiveness and truth. I assume that effectiveness and truth are orthogonal concepts. This means that effectiveness does not depend on truth, nor does truth depend on effectiveness. Effective models may rely on false content. And true content does not necessarily make a model effective. This assumption contrasts with alternative construals, such as that by Chang (2017). In his view, the truth of a proposition depends on its role in configuring a coherent activity (113). Coherent activities are, roughly speaking, activities whose inner operations (or subtasks) fit together harmoniously. Chang submits that coherent activities lead to success *ceteris paribus* (111). In this sense, Chang seems to suggest that a proposition is true if it is effective, i.e., its employment leads to reliable success. I depart from this and other similar views and hold effectiveness and truth to be independent.

Given that truth and effectiveness are independent, an important consequence of the effectiveness condition is that the content of effective models needs not be true about the target. Accordingly, an agent may have objectual understanding of a target without having knowledge of it. I consider this consequence to be correct. In fact, it is common to witness scientists and historians of science attributing understanding to other scientists who are able to use models successfully, even though the content of the model may be false. For example, consider the case of models that deploy principles of Newtonian mechanics. These models have been widely used – even today – with immense success: epistemic achievements have been attained, such as accurate predictions, and practical applications have been developed, such as controlled interventions into the dynamics of systems. However, since Einstein, physicists know that Newton's theory is strictly speaking false. Hence, models based on Newton's theory deploy false content. Still, it would be preposterous to deny (objectual) understanding to those scientists that used Newtonian models to develop successful applications.

Nonetheless, in some circumstances, scientists may expect their models to deploy true content. For example, consider the case in which scientists aim to use the content of a model as the basis of truthful representations of a target. This is a case in which success of the venture requires true content. That is, the effectiveness of a model coincides with its content being true. In this situation, we may attribute objectual understanding to an agent based on her possession of the object of ability, i.e., the representational model. But such attribution is justified by the model's effectiveness, not its truthfulness. In other words, the truth of the content, in this case, is a by-product of its effectiveness in being used for truthful representations.[72]

While truth may be a mere by-product of representational models that afford objectual understanding, it cannot be so construed in explanatory understanding. Explanatory understanding relies on model explanation. And false explanations are not explanations, as a matter of dictum.[73] Thus, there must be a truthfulness criterion that applies to model explanations as object of abilities in explanatory understanding. I proceed to discuss this criterion in more detail below, to which I refer as the "genuineness" criterion.

7.2.3 The Genuineness Condition on the Object of Abilities in EU

By definition, explanatory understanding relies on explanations. This is uncontroversial. But, as de Regt (2017) puts it, it involves "adequate" explanations (23). What is meant by "adequate" is the fountainhead of debates in the literature on explanatory understanding. My proposal – as expressed in EU – is that those adequate explanations (i.e.,

[72] Along these lines, Potochnik (2017) argues that "[t]ruth plays only a supporting role. When scientists do aim for truth, this is only because the circumstances make it so that truth is the best way to promote understanding" (103).

[73] Ylikoski & Kuorikoski (2010) analyze this dictum and adhere to it with some nuances: "The consensus in the theory of explanations seems to be that explanation is factive and that false explanations are only apparent explanations. However, it is not immediately clear what the rationale of this requirement is, and most arguments for this requirement seem little more than ideological exclamations: science simply *should* aim for true explanations. Moreover, truth is not an all-or-nothing affair, and it is widely accepted that intentional distortions of truth are sometimes needed when constructing explanations" (212). Having stated these nuances, Ylikoski and Kuoriokoski state: "We accept the basic realist view of the factivity of explanation because *totally false* explanations would provide incorrect answers to what-if questions and would thus provide only illusory understanding" (ibid).

those that afford the content of explanatory understanding) are genuine model explanations, regardless of their modal qualification. This position stands in between two poles concerning explanatory understanding, namely "factivist" and "non-factivist" approaches. I proceed to introduce these approaches.

Broadly speaking, factivist approaches to explanatory understanding submit that explanations that afford understanding must be true propositions, i.e., state the facts. This may be interpreted as saying that all propositions constituting an explanation must be true. However, this reading of the factivist stance is, in practice, exceedingly demanding. Scientists seem to deal with idealizations (i.e., falsehoods) regularly in explaining phenomena. Thus, the factivist stance can be relaxed in two steps. First, factivists may admit explanations whose propositions are true or approximately true. In this sense, some degree of distortion can be involved in what is stated in explanations. Second, factivists may admit that not all propositions in an explanation need to be true or approximately true. That is, some falsehoods in explanations are tolerable for understanding purposes. This relaxed version of factivism is often referred to as "moderate factivism" or "quasi-factivism."

In contrast, non-factivists argue that explanatory understanding can be obtained by means of explanations whose propositions, even central ones, are false. In fact, some philosophers would argue that some of our best scientific explanations are false (e.g., Cartwirght, 1983). This approach is *prima facie* more suitable to deal with the case of model explanations. Given that models contain idealizations, model explanations that refer to these idealizations contain propositions that are strictly speaking false about the target. These falsehoods may play a prominent role in the explanations. And yet, several philosophers argue that this sort of model explanations may provide explanatory understanding. Thus, these philosophers advocate a non-factivist approach to explanatory understanding.

Factivists and quasi-factivists are not persuaded by the non-factivist argument. They have their own strategies to deal with the presence of falsehoods in explanations that afford understanding. A first quasi-factivist strategy is to admit falsehoods in explanations as long as they are *innocuous*. For example, consider Strevens's account of understanding (2013), discussed above (see Chapter 6). Strevens's account can be characterized as a form of quasi-

factivism, in the sense of requiring that some propositions in explanations must be true, namely those doing the explanatory work. That is, propositions referring to the difference makers of the phenomenon must be factual. Yet, other propositions in explanations may be false, namely those that do not refer to the difference makers of the explanandum (i.e., non-explanatory propositions).

A second strategy is to admit falsehoods that enable explanatory understanding but are not themselves part of the content of such understanding. An example of this position is Lawler's "extraction view" of understanding (2019).[74] Her position is – strictly speaking – different from the quasi-factivist position in the following sense. In quasi-factivism, falsehoods are part of the *content* of explanatory understanding, as long as the falsehoods are not central (i.e., do not exert explanatory work as difference-makers). In Lawler's view, falsehoods are not part of the content of explanatory understanding. Rather, falsehoods are only enablers of explanatory understanding. In this sense, Lawler's view is strictly factivist insofar as the content of understanding is exclusively factive.

The latter does not mean that falsehoods play a merely heuristic role. Lawler argues that (felicitous) falsehoods are epistemically valuable, in the sense of containing elements of truth. Referencing Elgin (2017), Lawler states that falsehoods provide epistemic access to phenomena by exemplifying relevant features of them (24-5). Basically, Lawler's notion of felicitous falsehoods assumes that they can be decomposed into their false and true components. Thus, while the process of obtaining understanding relies on both components, the final outcome involves only the true component. The extraction view thus regards falsehoods as central to understanding, but not in terms of its content. Falsehoods are the raw material from which truths are extracted. Furthermore, falsehoods, in containing these elements of truth, are indispensable for understanding.

I proceed to assess the standing of EU in this debate among (quasi-)factivists and non-factivists. Against the non-factivist, I hold that false explanations are not explanations, as a

[74] Lawler introduces her account in the context of objectual understanding. Still, her views can easily be extrapolated to an explanatory account of understanding.

matter of dictum. Hence, false explanations cannot afford explanatory understanding. Accordingly, I construe EU as requiring genuine model explanations. This means that model explanations that afford understanding must be true about the target phenomenon. But they need not be actually true. Genuine model explanations may be roughly, partially, plausibly, or possibly true about the target phenomenon. In other words, genuine model explanations capture modal facts about the target.

Because genuine model explanations state facts about the target, even if modal ones, knowledge of their content and metacontent amounts to knowledge about the target. This contrasts with the case of effective models in OU, in which knowledge of the content of the model does not imply knowledge of the target. Recall that the content of an effective model had to be accepted as content of the target in order to meaningfully relate it to the target. In the case of genuine model explanations, the content of the model explanation is just content of the target. In this sense, acceptance is not required; mere knowledge does the work. Certainly, acceptance of genuine model explanations can still be adopted. That is, an agent might take a genuine model explanation as a basis for action (see discussion in the "downstream abilities" section below). But my point is that knowledge of genuine model explanations already affords a meaningful relation to the target.

Given that my account requires genuine model explanations to state facts – even if modal ones – it should be characterized as a factive account. However, most factivists seem to implement the term 'fact' in a narrower sense: facts *simpliciter* are the *actual* facts. And the term 'truth' – unqualified – is commonly used as *actual* truth.[75] This is a regrettable habit given that what is true is not limited to what is actual. Still, truth and actuality seem to be conflated in discussions on explanations. True explanations are meant as *actually* true explanations. That is, in traditional factivist approaches, understanding-affording explanations are those that refer to those factors that make an *actual* difference in the obtaining of the phenomenon.

[75] This may be a consequence of assuming a correspondence theory of truth. For example, consider the following characterization of the correspondence theory: "The basic idea of the correspondence theory is that what we believe or say is true if it corresponds to the way things *actually* are – to the facts" (Glanzberg, 2018; my emphasis)

This employment of the terms 'truth' and 'facts' as implicitly actual can be recognized in various factivist approaches to understanding. A good example of this can be noticed in Trout's analysis of the "sense of understanding" (2002). Trout argues that the sense of understanding that explanations often produce is not a reliable guide to truth. As Trout argues "this sense of satisfaction [i.e., the sense of understanding] is confidence that one enjoys an accurate description of the underlying causal factors sufficient (under the circumstances) to bring about the phenomenon we are examining. But confidence is, notoriously, not an indicator of truth" (214). This is certainly so. One may have counterfeit, as opposed to genuine, understanding. Trout argues that there are psychological biases that may cause this, such as hindsight or overconfidence. I do not reject this thesis. Instead, I just want to draw attention to what is meant by "genuine understanding."

Trout uses various historical cases to illustrate what he considers examples of the triumph of counterfeit understanding over genuine understanding, prompted by a mere sense of understanding. Consider the case of Ptolemy. Trout argues that Ptolemy's model of the solar system is based on false background assumptions and, accordingly, does not afford genuine understanding of the solar system. My reading is that Ptolemy was indeed wrong in believing that his claims concerning the solar system were actually true. Furthermore, Ptolemy was overconfident about his claims being actually true (Trout provides various quotes by Ptolemy that reflect this attitude). But this does not mean that his model could not be used to craft genuine model explanations, provided that an adequate modal qualifier was prefixed to them. More specifically, in light of the background knowledge and evidence available to Ptolemy, explanations derived from his model were plausibly true. In this sense, Ptolemy had genuine understanding of how-plausibly the solar system was organized.

Thus, I suggest that Trout is wrong in suggesting that Ptolemy's model does not provide genuine understanding simply because explanations derived from it are actually false. These explanations are actually false, but they are also possibly or even plausibly true (at least in Ptolemy's days). Ptolemy's mistake was overconfidence, not in claiming genuine understanding of the solar system, but in suggesting that such understanding was obtained via how-actually explanations. This lesson is expressed – in a more contemporary context –

by Parker (2014): "For instance, conceptual models and explanations developed by analyzing simulation results as surrogate observational data are treated as how-possibly or how-plausibly models and explanations, pending further empirical investigation; they are not, nor should they be, immediately accepted as how-actually models and explanations" (354). To put it bluntly, my disagreement with Trout amounts to the following: Trout suggests that a seemingly plausible explanation does not necessarily provide genuine understanding (in the "actualist" sense), and thus should not be confidently accepted.[76] My proposal is that a plausible explanation can be confidently accepted *as a* how-plausibly explanation. And, if so accepted, this how-plausibly explanation affords genuine explanatory understanding of what is plausible.

The implicit assumption that explanatory understanding requires actual explanations is not unique to Trout. For example, Strevens's account of explanatory understanding requires correct explanations, which may contain falsehoods, but their explanatory force relies on their *actually true* content. Strevens (2013) exemplifies this in the case of correct causal explanations: "On a causal account of explanation, for example, the internal condition [for an explanation to be a correct explanation] might stipulate that an explanation represent[s] a potential causal history for the explanandum; the external condition would then stipulate that this must be in addition the explanandum's *actual* causal history" (513; my emphasis).

Other factivists enforce actuality in explanations by means of representational accuracy.[77] Consider Salmon's approach (1998) to scientific understanding. After discussing the relations between explanations and understanding, Salmon concludes: "Understanding in the scientific sense involves the development of a world-picture, including knowledge of the basic mechanisms according to which it operates, that is based on *objective evidence* – one that we have good reason to suppose *actually represents, more or less accurately, the way the world is*" (90; my emphasis). If this is the scope of scientific understanding, then it seems that the contributions of non-actual explanations to understanding are null. After all, some

[76] "This subjective sense of understanding may be conveyed by a psychological impression that the explanatory mechanisms are transparent and coherent, or that the explanation seems plausible, and so should be confidently accepted" (ibid: 214).

[77] Rice (2018) claims that most accounts of how models explain phenomena require *accurate* representation of – at least – the difference-making factors (2796).

how-possibly explanations are not based on objective evidence (e.g., global how-possibly explanations). And some how-possibly explanations are – even knowingly – inaccurate representations of the way the world is (e.g., non-actual how-possibly explanations). I argue for the opposite. Inaccurate representations – and resulting false explanations – of the actual world may provide genuine understanding, insofar as the modal status of the explanation is adequately established. The same obtains for explanations that do not fare well in the light of available evidence.

Thus, I do not consider EU to be a factivist account of understanding in the traditional sense of relying solely on how-actually explanations. But this does not make EU a non-factivist account. My account still relies on explanations which convey facts, even if modal ones. In this sense, understanding-affording explanations may be false about what is actual, but then they must be true about what is plausible or possible. Genuineness is not limited to a certain modal status of the content of explanations.

Still, some non-factivist accounts are good allies of EU, namely those accounts that require understanding to be "true enough" as opposed to utterly true (e.g., Elgin, 2017; Potochnik, 2017). Consider Potochnik's account (2017). In her view, understanding is an epistemic achievement in the sense of being "subject to standards for success" (94). That is, genuine understanding requires "successful mastery, in some sense, of the target of understanding" (ibid). This consideration makes understanding an intersubjective phenomenon, because "successful mastery" is decided by collective standards. Potochnik, referencing Elgin, argues that truth is not the only – not even the best – standard to assign success in understanding. Rather, something lesser is epistemically acceptable to attribute success, namely "true enough" representations. In this sense, Potochnik's account embodies an inversion of the traditional view of the aims of science: "Truth plays only a supporting role. When scientists do aim for truth, this is only because the circumstances make it so that truth is the best way to promote understanding" (103).

My account does not renounce to truth altogether as a standard for success. But it does consider a degree of approximation to what is actually true. In this sense, understanding-affording explanations are "actually-true enough." If they fall short of being actually true, they

must be true according to another modal qualification, e.g., plausibly or possibly true. In this sense, while it is the case that scientists may aim for a specific modal status of explanations, this does not mean that other modalities of explanations do not provide any understanding at all. For example, factivists traditionally aim for how-actually explanations (as argued above). It is in light of this desideratum that non-actual how-possibly explanations are deemed incorrect and thus do not afford understanding. But that, I suggest, is an inadequate construal of explanatory understanding in actual scientific practice. Non-actual how-possibly explanations provide *some* understanding. That is, there are different forms of modal understanding associated with the different modal qualifications of understanding-affording explanations. The modal status of genuine model explanations does not decide the obtaining of explanatory understanding but rather only its quality. In other words, the modality of explanations is a dimension of assessment of understanding.

Other philosophers have argued that understanding is a modal notion, although this is spelled out differently. These accounts provide some support to my views. Thus, I consider it relevant to briefly review some of them, namely those due to Lipton (2009), Le Bihan (2017), and Reutlinger *et al.* (2018). First, Lipton (2009) argues that non-actual how-possibly explanations may afford genuine understanding.[78] Hinting at Descartes, Lipton says: "We may improve our understanding as to why a phenomenon is as it is not just by being given a more or less accurate account of how it actually came about, but by being given even a wildly divergent account of how it could have come about" (50). Lipton concludes that "[a]ctual [i.e., genuine] understanding may arise from being shown how something is possible" (ibid). In particular, this can be conveniently spelled out in a causal-mechanistic approach to explanation: "Information about a possible mechanism may give oblique information about the actual mechanism (telling us something about a class into which it falls) by describing a mechanism in the same class, namely the class of mechanism that can generate this [kind of] phenomenon" (51).[79]

[78] Lipton (2009) refers to non-actual how-possibly explanations as "merely possible" explanations (51).
[79] Batterman and Rice (2014) express a similar insight with their notion of "minimal model explanations," in which universality classes responsible for a same kind of phenomenon are established.

Interestingly, Lipton does not seem to consider the understanding derived from non-actual how-possibly explanations a form of explanatory understanding: he discusses this as a case of "understanding without explanation."[80] Arguably, this is due to the fact that most theories of explanation take non-actual explanations to be no explanations at all because non-actual explanations are known to be false about what is the case (e.g., "The consensus in the theory of explanation seems to be that explanation is factive and that false explanations are only apparent explanations"; Ylikoski & Kuorikoski, 2010: 212). Still, non-actual explanations may be true about what is plausibly or possibly the case. In this sense, non-actual explanations may be genuine explanations, provided that they have an adequate modal qualifier. Thus, I suggest that the understanding afforded by non-actual how-possibly explanations should be counted as explanatory understanding, contra Lipton.

Second, Le Bihan (2017) proposes an account of modal understanding, which consists in the ability of navigating the various possible worlds in which phenomena arise (112). In other words, understanding a phenomenon involves not only knowing how it actually happened, but also knowing how it could possibly have occurred. She crafts this view by studying theories and associated models that knowingly misrepresent the way the world *actually* is. In this sense, Le Bihan's modal understanding is an ability to craft explanations from an existing model (cf. "ability endogenous to the object of ability" above). In particular, the exercise of this ability results in explanations of variable modal status, from how-possibly to how-actually explanations.

Third, Reutlinger *et al.* (2018) explicitly discuss the role of how-possibly and how-actually explanations in achieving understanding. Their account is basically a refinement of Strevens's views and it is intended to account for the understanding gained with toy models. Given that the BK and OFC models can be characterized as toy models (highly simplified models), Reutlinger *et al.*'s account is particularly relevant to my argument. Thus, I take some space to elucidate their views.

[80] To be clear, I agree with Lipton on the claim that there are other non-explanatory forms of understanding. In particular, OU is one example. However, I suggest that the understanding derived from non-actual how-possibly explanation is explanatory understanding, assuming a modal approach to explanation.

Strevens's so called "simple view" is expressed as follows:

(SV) An individual scientist S understands phenomenon P via model M iff model M explains P and s grasps M.

Reutlinger *et al.* refine SV in four ways. First, they assume a naturalistic approach towards grasping. It is a scientific matter to figure out what grasping is, not a philosophical one. Second, understanding is contextual. Roughly, understanding depends on the understander and her community. Third, explanations may have different modalities, in particular how-possibly and how-actually explanations. The understanding resulting from these different modalities of explanations is also distinct in its modal status, namely how-possibly and how-actually understanding. Fourth, Reutlinger *et al.* assume neutrality concerning the kind of explanations used in affording explanatory understanding. That is, all accounts of explanation stand, in principle, on equal grounds.

Thus, Reutlinger *et al.* deliver a refined simple view:

(RSV) Scientist S understands phenomenon P via model M in context C if and only if one of the following conditions holds:
1. Scientist S has a *how-actually understanding* of phenomenon P via model M in context C if and only if model M provides a how-actually explanation of P and S grasps M.
2. Scientist S has a *how-possibly understanding* of phenomenon P via model M in context C if and only if model M provides a how-possibly explanation of P and S grasps M.

While I endorse most of RSV, I consider their distinction between how-actually and how-possibly understanding misleading. How-actually and how-possibly explanations provide a different content to our understanding of an explanandum phenomenon. It is in this sense that the distinction is adequate. However, the distinction may be confounded as a distinction in the modal status of the attained understanding. This is not the case, or so I argue. The understanding attained via genuine how-possibly and how-actually explanations has the same modal status: actual (or genuine) understanding of the explanandum phenomenon. As Lipton expresses (2009), how-possibly explanations do not merely provide knowledge about

possible worlds. They provide knowledge about the actual world. More explicitly, how-possibly explanations provide knowledge about an actual phenomenon by indicating possible ways to generate the actual phenomenon (ibid: 51).

In addition, RSV could easily be extended to other modal qualifications. For example, consider the epistemic modality of how-plausibly: scientist S has a *how-plausibly understanding* of phenomenon P via model M in context C iff model M provides a how-plausibly explanation of P and S grasps M. And so on for other modalities. Along these lines, I suggest that the modal qualifiers attached to explanations should be taken as values along a dimension of assessment of the resulting understanding. More explicitly, it is not the case that some modalities afford understanding and others do not. Rather, explanations with different modalities afford different content to understanding. How to evaluate such content in terms of success is a different, more contextual, question, which I address below.

With this, I finish the section on the object of abilities in EU and OU. On reflection, I consider that my views on scientific understanding provide a reasonable compromise between factivist and non-factivist accounts in two senses. First, EU is – strictly speaking – a factivist account, but it incorporates a modal dimension. Thus, explanations that are false about what is actual may still provide understanding by being true about what is plausible or possible. Second, while EU is factive, OU needs not be. Factivity in OU is only a side-effect of the fulfilment of the effectiveness condition.

To close this section, I recall some of the main conclusions withdrawn:

i. The object of ability in OU amounts to the active content of a representational model plus its metacontent.
ii. The object of ability in EU amounts to the active content of a model explanation plus its metacontent.
iii. The object of abilities in OU must be effective in order to afford understanding.
iv. The object of abilities in EU must be genuine in order to afford understanding.

v. An agent understands a target explanatorily if and only if she possesses the object of ability in EU, i.e., she explicitly knows the active content of a genuine model explanation, its metacontent, and has modal knowledge about them.

vi. An agent understands objectually if she possesses the object of ability in EU, i.e., she explicitly knows the active content of the model, the metacontent, and has modal knowledge concerning them. In addition, she must accept the model's content as content of the target, allowing her to use the content for epistemic achievement and practical applications concerning the target.

7.3 The Abilities of Scientific Understanding

While the objects of abilities are concerned with the content of understanding, the abilities of understanding are concerned with those competent actions that relate to this content. The relations of abilities to the content of understanding may be described in terms of generating the content or capitalizing on it. Accordingly, I suggest that abilities can be grouped in two main categories, namely "upstream" and "downstream" abilities (Figure 7-1). Upstream abilities are those abilities involved in the crafting of representational models and model explanations. Downstream abilities are those abilities that are enabled by having representational models and model explanations. Note that, for the purposes of my account, I do not distinguish abilities deployed in EU from those deployed in OU. After all, in both cases, the abilities are construed as actions related to content. Whether this content resides in a representational model or a model explanation is not substantial to my discussion of abilities.

I suggest that upstream and downstream abilities are each sufficient to attribute ability to an agent. This is why, in both EU and OU, I state that crafting *and/or* using the object of ability must obtain. Surely, a state of understanding can be assessed as higher or lower depending on whether it contains both upstream and downstream abilities or only one of them. But this is a different matter: the assessment of a state of understanding should not be confounded with the attribution of understanding to an agent. The attribution of understanding, based on ability, follows the criteria stated above: abilities are a necessary component of EU and a sufficient one in OU.

The relation that abilities have to their objects (i.e., the content of models and model explanations) is worth a few paragraphs of elaboration. In the case of upstream abilities, scientists, by constructing models, are generating content. That is, there are propositions that are now true about a newly existing entity, namely the model. However, scientists do not need to be explicitly aware of this content. And even if they were explicitly aware of it, they do not need to explicitly know its metacontent. In other words, scientists may be able to craft models without being able to explicate them. Thus, there are cases in which upstream abilities are deployed, but possession of their objects – i.e., an explicitly known content and metacontent – is lacking.

Similarly, in the case of downstream abilities, scientists can relate to the content of models by using the models in successful applications. However, the successful employment of models does not necessarily require explicit knowledge of the active content nor its metacontent. In other words, there are cases in which downstream abilities are exhibited without possession of the object of abilities. Using a familiar term in the literature, there are cases in which downstream abilities involve tacit knowledge about the model. The lesson is the following: scientists may be able to make explicit the content that their abilities generate or capitalize on, but this is not necessary to execute abilities successfully.

However, I suggest that there are important exceptions to this generalization. Thus, before addressing the topic of upstream and downstream abilities in more detail, I consider it relevant to have a preliminary discussion on the topic of tacit knowledge. In this section, it becomes clearer why possession of the object of abilities in EU is necessary, as opposed to merely sufficient in OU.

7.3.1 Abilities and Tacit Knowledge about Models

Polanyi (2009) develops an account of tacit knowledge and scientific discovery. This account is based on a notion borrowed from Gestalt psychology, viz. the notion of "subception". The slogan that captures the conceptual role of 'subception' is that "we can know more than we can tell" (ibid: 4). In particular, Polanyi suggests that there is a "tacit power" – or, more precisely, a skill – in the way we shape and integrate our experiences. It is

through the exercise of this skill – an instance of "know-how" – that explicit propositional knowledge is possible. In this sense, Polanyi suggests that "knowing-how" and "knowing-what" are two sides of the same coin.

I have some reservations about Polanyi's overall approach to tacit knowledge, as I expressed in Chapter 3. In particular, the "personal" character of tacit knowing makes it somehow obscure. However, there are resources in his account that are expedient for explicating how tacit knowledge is involved in the execution of abilities that relate to models in scientific understanding. Thus, I suggest bracketing what is meant by "personal" and simply resort to his proposed structure of tacit knowledge. (Collins (2010) replaces the "personal" with the "social", making tacit knowledge less mysterious, although not less complex.)

According to Polanyi, tacit knowledge involves two related components, which can be distinguished in terms of attention. On the one hand, there is a distal component to which we attend. The distal component is specifiable. That is, we know the distal component explicitly. On the other hand, there is a proximal component that we are aware of but only through our attending to the distal component. In this sense, the proximal term is tacit. This binary structure formed by the proximal and the distal components embodies a *from-to* structure. That is, we attend *from* the proximal component *to* the distal component.

As an illustration, Polanyi discusses the ability to recognize a face. Most people have the ability to recognize familiar faces, although how such ability is exercised can hardly be stated explicitly. In this sense, as Polanyi argues, the ability to recognize a face from particular features is a tacit form of knowledge: we may know a physiognomy without being able to specify the particulars that allow us to recognize the physiognomy. More precisely, we are not explicitly aware of how the integration of particular features allows us to deliver a recognition of the whole, i.e., the familiar face. In this example, the face is the distal component to which our attention is directed. The particular features of the face are the proximal component. We are typically aware of the particular features of the face but only in their allowing us to recognize the face. That is, we are aware of the proximal component (i.e., particular facial features) in the appearance of the distal component (i.e., the face as a whole). The knowing-how in this case is the ability of tacitly integrating particular features (the

proximal component) into a whole (the distal component). And the knowing-what is the specifiable whole – i.e., the face – to which we attend from the particular features.

It is tempting to use a similar analysis in the case of scientists' relation to models. Scientists may be able to recognize a model. This could be the case even if scientists are unable to specify which aspects of the model are those that identify it. In this example, the model is the distal component, to which attention is directed: the model is being recognized. The content and metacontent of the model is the proximal component. This analysis seems plausible, although not especially valuable: the recognition of models is not a particularly praiseworthy ability in science.

What is particularly important in science is the meaning that a model has for a certain venture. And Polanyi has some insights that contribute to this topic. To account for meaning, Polanyi introduces two relations that are embodied in the from-to structure of tacit knowledge. First, there is a "functional" relation between the proximal and distal components. That is, we know the proximal component only by relying on our awareness of it for attending to the distal component (ibid: 10). In other words, while we know the whole explicitly, we only know the particulars tacitly. Second, there is a "phenomenological" relation between the proximal and distal components. That is, we are aware of the proximal component (i.e., that from which we are attending) in the appearance of the distal component (i.e., that to which we are attending) (ibid: 11).

The functional and phenomenological relations between the proximal and distal components enable the emergence of meaning. This is the so-called "semantic" aspect of tacit knowledge. Meaningless proximal components become meaningful as they are used to know the distal component. In this sense, meaning could be said to be "away from ourselves," i.e., in the domain of the distal component (ibid: 13). Polanyi describes the process by which proximal components acquire meaning as an "interpretative effort" (ibid: 12).

At this point, Polanyi introduces the notion of understanding. He suggests identifying "understanding" with his notion of tacit knowledge: "Since tacit knowing establishes a meaningful relation between two terms, we may identify it with the *understanding* of the

comprehensive entity which these two terms jointly constitute. Thus, the proximal term represents the particulars of this entity, and we can say, accordingly, that we comprehend the entity by relying on our awareness of its particulars for attending to their joint meaning" (ibid: 13). Understanding is thus explicated as the meaningful relation between proximal and distal components of tacit knowledge. In other words, understanding is the outcome of Polanyi's interpretative effort. In this explication, the object of understanding is not the distal, explicitly known, component. Rather, the object of understanding is the comprehensive entity constituted by the meaningful relation established between proximal and distal components.

To keep things clear, Polanyi's notion of understanding is not the same as my notion of scientific understanding, neither in explanatory understanding nor objectual understanding. What I am suggesting is that his notion of understanding – as tacit knowledge of a comprehensive entity – plays a role in accounting for the meaningful relation that scientists have towards models in the execution of abilities. Models are a meaningful product of upstream abilities. And they are a meaningful resource for downstream abilities. This meaning emerges in the interaction between proximal and distal components, i.e., between active content plus its metacontent and the model as a whole. This meaningfulness does not require direct attention to the model's active content plus its metacontent. That is, explicit knowledge of the active content plus its metacontent is not required for meaningfully relating to a model. For example, a scientist might be able to recognize a model's effectiveness – and use it successfully – without being able to satisfactorily explicate the specific content and metacontent that makes it effective.

To be clear, this does not mean that explicit knowledge of the active content and its metacontent is not possible. Certainly, scientists can investigate their models – focus on the proximal component – and discover (i.e., explicitly know) their content and metacontent. However, this might come at a cost. Polanyi suggests that, in some cases, explicit knowledge of the proximal component disturbs the meaning of the comprehensive entity. (The best practical illustration for this comes by noting that focused attention on the movement of our fingers disturbs the meaning of our performance in playing the piano.) In any case, even if explicit knowledge of the active content of models and their metacontent is possible, the point is that it is not necessary for the successful execution of abilities that relate to models.

There is an obvious exception to the latter point, namely those abilities which – themselves – consist in making explicit a piece of a model's content and its metacontent. A notable example of this is the ability of explaining with models. In order to elucidate this point, it is worth recalling the epistemic view on explanations (Chapter 4). The epistemic view states that *explanations are activities* that an agent conducts in order to increase understanding and knowledge of a phenomenon (Wright, 2012: 382). Explanations may be *embodied* in particular acts of communication, in texts that contain the explanatory information (including models) or in mental representations (cf. Craver, 2014: 35). But, in order to say that these are embodiments of explanations, the explanatory content must be made explicit in them and presented as an explanation. To put it bluntly: there is no such thing as explaining tacitly. Pointing at a model with an adequate – but tacit – explanatory content does not amount to explaining.

In arguing this, I assume that a model, as a sort of text, does not explain. A model is not, itself, the ability of explaining. In other words, models are not intrinsically explanatory. Explaining is an activity that human agents do, often by making explicit the explanatory content of models. To be fair, this claim contrasts with other approaches in the literature. In this regard, it is worth quoting Rohwer & Rice (2016) at length, who discuss one of these approaches: "The easiest relationship to identify between models and explanations is a simple *identity relationship* between all the propositions of an explanation and propositions within a model. And many have claimed that a given model just is an explanation […] In fact, sometimes "explanation" and "model" are used interchangeably and there are numerous accounts of how scientific models explain in the literature. According to our framework, a model will be an explanation only if it includes a set of propositions sufficient to explain the target explanandum" (1132-3; my emphasis).[81] Giving credit to Rohwer & Rice, I admit that models are often referred to as explanations, and often said to be explanatory. But, given my argument above, such characterizations seem hyperbolic or elliptical. A model may "include

[81] Rohwer and Rice (2016) have a more nuanced stance on this issue. They consider this approach admissible but not exhaustive of the relations between models and explanations. In fact, in a more recent paper, they discuss a position more aligned with mine, in which a "model is used to discover and construct an explanation, but it may not *be* an explanation on its own" (Rice *et al.* 2018: 2)

a set of propositions sufficient to explain the target explanandum", i.e., have explanatory content. But, unless the explanatory content is made explicit and presented as an explanation, I do not consider it reasonable to say that the model in question explains. In sum, the success of the ability of explaining with a model relies on revealing the content to which it relates, i.e., making explicit the content of the model explanation.[82]

So much for "explaining". My concern is a more general one, namely the role of explicit knowledge in abilities, particularly those related to EU. As part of my account of EU, I submit that the possession of the object of abilities is necessary. That is, if an agent has EU, then she possesses explicit knowledge of the content of a genuine model explanation, its metacontent, and modal knowledge about them. I intend to show that this more general claim is consistent with the rest of my account, thus making explicit knowledge necessary *in the context of my account*. (I cannot prove that explicit knowledge is a necessary requirement in any construal of explanatory understanding. I can only provide what I consider agreeable reasons for this – as those given above in the context of the ability of explaining – and show that the requirement hangs together with the rest of my account.) I consider two exhaustive scenarios: one in which abilities of EU are not possessed, and another in which abilities of EU are possessed. In the first scenario, the argument is straightforward: given that EU is a species of OU, if there is no possession of abilities of EU, then there must be possession of their object. This is by definition: at least one component (ability and/or its object) must be present in OU.

The second scenario is more problematic. The idea is to argue that, even in cases where abilities of EU are possessed, their objects must be possessed as well. Here, I turn to the *externalist* scope of my account. The idea is to establish criteria that allow an external agent to attribute EU to another agent, as opposed to merely OU. In other words, an external agent must be able to assess whether the object of the ability is a model explanation, as opposed to the whole of a model. This matter is not transparent from the mere inspection of the ability. There are several abilities whose object may be, as a matter of contingent fact, a model explanation, such as predictions, manipulations, classifications, to name a few. But, as

[82] In Chapter 6, I argued that surrogative reasoning, in the form of model explanation, required knowledge about the causal or mathematical structure of a model and imputation to its target. Now, I can emphasize that such knowledge must be explicit.

Lipton (2009) argues, there are also non-explanatory paths to execute these abilities. Thus, witnessing an agent engaging in these tasks may suffice to attribute understanding (this is Lipton's conclusion), but such attributed understanding is objectual. Additional evidence is required to attribute explanatory understanding; to say that these abilities relate to explanatory knowledge. In this sense, the problem is a subtle one. It is not that the abilities of EU require possession of their object. They can be executed based on tacit knowledge of model explanations. Rather, the problem is that EU cannot be attributed if such knowledge remains tacit.

Thus, I conclude my discussion of the relations between abilities and tacit knowledge. I have argued that abilities that relate to models and model explanations may be executed with tacit knowledge of their content. However, in order to attribute EU, possession of the object of ability is required. Having clarified the role of tacit and explicit knowledge in the exercise of abilities, I proceed now to discuss the two varieties of abilities of understanding, namely upstream and downstream abilities.

7.3.2 Upstream Abilities (or the Understanding of Theories)

The crafting of representational models and model explanations is a skillful task, an ability. This craft – modeling – implicitly generates content, namely the content of the representational model and the content of model explanations. In other words, the crafting of models supervenes on the generation of content. However, this content does not need to be explicitly known, as discussed above.

Modeling can be analyzed in terms of more specific abilities involved in it, i.e., upstream abilities. Several, and often intertwined, upstream abilities are exerted in modeling. Each upstream ability can be construed as the ability to use resources of a certain sort in constructing models. I do not intend to provide an exhaustive examination of the various resources that are used in the construction of models. Instead, I focus on one general resource which has a broad scope and is often discussed in the literature as a prime one in the construction of models, namely theories.

There is a large, and intensely debated, body of literature on scientific theories and their relations to models. A survey and discussion of this literature goes beyond the scope of this work (for an overview, see Portides, 2017). My strategy is to assume a well-established stance on the nature of theories and their relations to models and elaborate on it: the "models as mediators" stance. This stance is well described by exponents such as Morgan & Morrison (1999), Giere (2006) and more recently adopted by de Regt (2017) in the context of scientific understanding with models. According to this position, a theory is a collection of principles – often referred to as "theoretical principles" – which provides the basic tools for the construction of models (de Regt, 2017: 31-2; cf. Giere, 2006: 60-2). This position conveys an asymmetric relation between theories and models of a genetic sort. Accordingly, models are conceived as distinct from the theories they embody. Furthermore, once constructed, models enjoy a certain degree of autonomy from the theories that are utilized in their construction.

The advantage of using this approach to theories is that the notion of theory is minimal enough to cover various kinds of resources utilized in the construction of models. After all, a great variety of entities fall into the description of "principles which provide the basic tools for the construction of models." For example, theoretical principles may include concepts, calculi, hypotheses, etc. In this sense, theories are the toolbox from which resources are taken to construct models (cf. "theory as tool" in Suarez & Cartwright, 2008: 62-3). The ability to use theories in the construction of models is, itself, a set of upstream abilities, each one of them consisting in the ability to use a resource of a particular sort to be implemented in the models.

The ability to use theories to craft models has been examined thoroughly by de Regt (2017; see also de Regt & Dieks, 2005). I rely mostly on his views but refine a few aspects in light of the lessons learnt in the context of my case studies. De Regt (2017) distinguishes between "understanding of phenomena" (which he construes as a form of explanatory understanding) and "understanding of theories." His criterion for understanding phenomena is: "A phenomenon P is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency" (92). A theory is intelligible if it exhibits (clusters of)

qualities – in one or more of its representations – that are valued by the scientists in the sense of facilitating the use of the theory (40). In this sense, intelligibility is a sort of measure of the fruitfulness of a theory for constructing models, which is highly contextual. Then, understanding of theories amounts to the ability of choosing and using theories to build models and model explanations, based on skills and judgments of the scientists in a given context.[83]

I suggest three refinements to de Regt's approach to the ability of using theories in the construction of models and its role in scientific understanding. First, I emphasize an often-overlooked sense in which theories are employed in the construction of models. Theories not only play a role in the construction of models from zero, i.e., in the construction of new vehicles. Theories are used in the interpretation (and re-interpretation) of already existing vehicles. This interpretational task amounts to constructing a model because models are interpreted vehicles (see Chapter 5). In other words, a new interpretation of a same vehicle generates new content and thus amounts to the construction of a new model.

Second, theories used in the construction of models need not comply with "the basic epistemic values of empirical adequacy and internal consistency," as de Regt suggests. Certainly, these values may be common desiderata, but not necessary for understanding. In fact, de Regt admits that these values may experience contextual trade-offs (93). However, he ultimately proposes that they appear to be a necessary condition for an explanation to produce understanding (ibid). His intention in demanding these values is to distinguish scientific theories from pseudo-scientific ones, such as astrological theories. I estimate that this is the wrong criterion because scientific theories may be empirically inadequate and internally inconsistent, even if provisionally. And pseudo-scientific theories, like astrological ones, may be coherent and empirically adequate, even if by accident. According to my account of scientific understanding with models, the burden is not on the theories, but on the models. In OU, the models must be effective. In EU, the model explanations must be genuine. Whatever theories lead to the construction of models with these qualities are acceptable.

[83] I am aware that explicating abilities in terms of other abilities seems to lead to an infinite regress. The hope is that a theoretical resource, along the lines of Chang's "operational coherence" (2017), may provide a solution to this problem. I do not solve this problem here.

Note that, by following these standards, pseudo-scientific theories are also dealt with. Pseudo-scientific theories do not lead to effective models in the sense that the resulting models do not lead to success *reliably* (see 6.2.2 above).

There is a particular reason why I do not demand theories to be empirically adequate and internally consistent. What I have in mind are those theories that are used in the construction and interpretation of models in *exploratory stages* of research. I explore this topic further in the next chapter. However, a few ideas are worth discussing here. My suggestion, contra de Regt, is that theories that are empirically inadequate and are not fully internally consistent may still be used to advance understanding, whether as OU or EU. In fact, empirically inadequate and internally inconsistent theories are often used in exploratory stages of research. This is admissible provided that the resulting models are effective or lead to genuine model explanations. Note that genuine model explanations may be obtained from using empirically inadequate theories. This is because genuine model explanations may be how-plausibly and how-possibly explanations, which do not need to comply with the actual evidence (i.e., empirically inadequate).

The third refinement is concerned with de Regt's criteria for intelligibility of theories. He submits that the understanding of theories – the ability to use them – is closely related to their intelligibility. He provides the following criterion for the intelligibility of theories (CIT): a scientific theory T is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations (ibid: 102). De Regt provides various cases to illustrate the operation of CIT. Although not mistaken, CIT is notably incomplete. In fact, de Regt himself acknowledges that CIT does not exhaust the criteria for the intelligibility of theories.

In particular, CIT does not fare well in the context of my case studies. Consider the BK model and the theoretical principles used in its construction. BK were not able to recognize qualitatively characteristic consequences of these principles without the aid of simulations. In this sense, BK's efforts can be characterized as experiments with a real component of surprise. Certainly, BK attempted to simulate the qualitative features of earthquake behavior with their model, but the point is that this was not foreseeable from the principles they

assumed in the construction of their model. To make a broader case, CIT seems to fare poorly in the context of understanding emergent phenomena in complex systems via computer simulations.

It is interesting to note that de Regt tests CIT in a case of computer simulations of complex systems in meteorology. He argues that, even in this case, meteorologists are not merely occupied with making correct predictions via the "brute force" of computer simulations (ibid: 106). De Regt suggests that meteorologists also aim for formulating intelligible theories. Otherwise (the argument goes), they would fail to understand the weather. Certainly, it might be a common desideratum to formulate theories in a way that allows scientists to recognize their qualitative consequences without performing exact calculations, even in meteorology. However, this is a desideratum that scientists, and in particular meteorologists, often do without. Scientists do run simulations to appreciate the consequences of their theories, even qualitative ones.

A relevant case is that of "weakly emergent" phenomena. Bedau (1997) defines weak emergence as it follows: a macrostate P of a system S with microdynamic D is weakly emergent iff P can be derived from D and S's external conditions but *only by simulation* (378; my emphasis). In the case of weakly emergent phenomena, a theory may be available that inform us about the microdynamics of a system. But knowledge of this theory, by itself, does not allow scientists to recognize its qualitative macroscopic consequence without the aid of a computer simulation. In this sense, by following CIT, scientists fail to understand weakly emergent phenomena, whether in meteorology or other disciplines.[84]

This suggests that CIT is an inadequate criterion in some contexts, especially in the case of simulations of complex phenomena. Indeed, it seems an exceedingly demanding criterion for the understanding of theories to ask scientists to foresee the qualitative features of implementing a theory in a model that involves complex relations. As an alternative, I

[84] Arguably, one reason why CIT does not fare well in the context of theories implemented in computer simulations of complex phenomena is that most of de Regt's case studies are cases in 19th and early 20th century science. There are, to be clear, citations to contemporary science in his book. But de Regt's core source of inspiration lies in historical cases, such as theories due to Newton, Maxwell, Boltzmann, Huygens, Boyle, and the like.

suggest introducing an additional criterion for the intelligibility of theories. Call it CIT_2: a scientific theory T is intelligible for scientists (in context C) if they can implement T in an effective (computer) simulation.[85]

In sum, I have introduced three refinements to de Regt's notion of the understanding of theories, conceived as the ability to use theories in the construction of models. First, different theories may be used to reinterpret existing vehicles. Second, theories that inform the construction of models need not be empirically adequate nor internally consistent, especially in exploratory stages of research. Third, an alternative criterion for the intelligibility of theories is introduced which captures the role of theories in simulations of complex phenomena. Now, I proceed to elaborate on the understanding of models and distinguish it from the understanding of theories.

7.3.3 Downstream Abilities (or the Understanding of Models)

Downstream abilities of understanding are the various skillful uses that scientists give to models in general and to model explanations in particular. These abilities rely on both the intrinsic features of models and model explanations – namely their content – as well as on the skills and interests of scientists in using the explanations. Foreseeably, there is a great variety of uses that can be given to models and model explanations. I do not attempt to provide an exhaustive review of these uses. Rather, I have two purposes in mind for this section. First, I discuss some general issues concerning the intelligibility of models. Second, I provide some examples of downstream abilities that may shed some light upon the kinds of achievements that scientific understanding with models encompasses.

[85] In the case of explaining complex phenomena with models, simulations seem to be explanatory in hindsight, i.e., once the results of the simulations are obtained. Some philosophers have criticized the attribution of genuine explanatory understanding to that afforded by explanations in hindsight (e.g., Trout, 2002). Explanations in hindsight, the argument goes, provide a sense of understanding which is not genuine understanding. While there is some merit to this argument, it seems to preclude genuine understanding from scientific practice that involve simulations of complex phenomena. I would argue that hindsight itself is not the problem, but rather the lack of awareness of its effects (the so-called "hindsight bias" or "I-knew-it-all-along" effect). I would argue that this is not a serious threat in serious scientific research. Scientists, for the most part, are aware of the effects of hindsight. In this sense, the results of simulations of complex phenomena are not posed as "expectable," even though they can be – on reflection – explained by considerations implemented in the simulations.

The ability to use models can be treated analogously to the ability to use theories above. This means that the ability to use models can be reworded as the "understanding of models" (in analogy to the "understanding of theories"). And the understanding of models can be explicated in terms of their being intelligible. The intelligibility of a model is the perceived value of a model and its qualities in terms of facilitating its employment (in analogy to the "intelligibility of theories"). As a result, the understanding of models – like the understanding of theories – is highly contextual: the qualities of models are differently appreciated in distinct contexts.

In spite of the analogies, the understanding of theories and the understanding of models are not equivalent, nor can the understanding of models be reduced to the understanding of theories. The ability to use models transcends the understanding of the theories employed in their construction. More explicitly, there are resources in a model that are unique to them, not found in the theories used for its construction. The understanding of models involves relating to these unique resources. As an example, consider the novel resources provided by simulations, namely their outputs. Simulations are a sort of experiment in which scientists acquire novel results – even get surprised by them – and learn aspects about the workings of the model. These discoveries and lessons are not an intrinsic part of the understanding of theories used in the construction of models. This is notably the case in computer simulations of complex phenomena (e.g., see the discussion on "weak emergence" above). Thus, the understanding of models is distinct from the understanding of theories.

As I argued above, the ability to use models may rely on tacit knowledge of their content, i.e., without possession of the object of ability. In this context, a model acquires meaning as a comprehensive entity. Part of this meaning includes the value that scientists attribute to the model – as a whole – in terms of its employment, i.e., its intelligibility. But in other cases, scientists use models by focusing – explicitly – on specific aspects of their content and metacontent, i.e., by possessing the object of ability. In this case, intelligibility refers to

the value attributed to those specific aspects that make the model useful. One could think of this as a distinction between a coarse-grained and fine-grained intelligibility of the model.[86]

The intelligibility of a model – both coarse and fine-grained – is closely related to the non-representational interpretation of its vehicle. After all, the qualities of a model are the qualities of its vehicle as conceptualized and schematized in a non-representational interpretation (see Chapter 5). In fine-grained intelligibility, the attention is on the qualities of the model, as conceptualized and schematized in the non-representational interpretation. In coarse-grained intelligibility, the attention is on the whole model, while the conceptualization and schematization become tacit. Given that the non-representational interpretation of a vehicle is based on scientists' commitments, it follows that the so-interpreted qualities are bestowed to a model in a way that promotes its employment. In other words, models *are made intelligible* by means of the very non-representational interpretations that brings them about. As a caveat, the intelligibility of a model does not secure its effectivity. That is, models may be valued as useful – as a whole or for some of their features – and still may not lead to reliable success. A main reason for this is that, even if a model is interpreted in ways that promote its use, there are objective constraints in its performance set by the nature of the vehicle. Thus, intelligibility is not a success term.

After discussing some general issues on the intelligibility of models, I proceed to deliver an overview of downstream abilities. My intention is not to provide an exhaustive review of downstream abilities. Rather, I intend to give a taste of the variety of downstream abilities. I briefly discuss these abilities. I do not follow a particular scheme for this overview, although some proposals are available in the literature (e.g., see "intrinsic" and "extrinsic" usages in Frigg, 2003: 626). My impression is that these proposals fail to convey the intricate intertwining of downstream abilities. Thus, while I intend to distinguish downstream abilities conceptually, I do not consider that such distinction is straightforwardly attainable in practice.

[86] The same distinction can be applied to theories, where fine-grained intelligibility is focused on specific principles that conform a theory.

To begin with, a relevant use of models – one which plays an important part in my account of understanding – is *explaining phenomena with models*. "Explaining" is construed as a downstream ability under two assumptions. First, there are upstream abilities that lead to the crafting of model explanations, generating explanatory content. Second, there are activities endogenous to the object of ability that consist in the assemblage of the active content and management of its metacontent. Thus, explaining is a downstream ability in the sense that its object already exists, and it is possessed.

As I argued above, explaining with models involves possession of a genuine model explanation. But intact possession of such object does not amount to explaining. Explaining involves using that object adequately in the right context. An agent must be able to recognize those contexts in which the content of a model explanation is explanatory. In Chapter 6, I argued that part of this ability relies on an assessment of resemblance between model and target explanandum. This assessment of resemblance is conducted between the behavior of the target and the outputs of a model. If resemblance is attributed, then the causal/mathematical structure of the model (i.e., content and metacontent) may be imputed to the target as a model explanation, with a modal qualifier to be specified. Thus, explaining implies EU: it requires possession of a model explanation and the ability to use it.

Because of this, the insights of pragmatic approaches to explanation seem relevant to explanatory understanding. From a pragmatic perspective, explanations are not merely pieces of content to be possessed. Rather, explanations are knowledge to be used, e.g., as answers to why-questions (van Fraassen, 1980) or solutions to problems (Mantzavinos, 2016). Along the same lines, EU also requires the ability to act upon model explanations.[87] Thus, pragmatic accounts of explanation are closely related to what I deem to be explanatory understanding. Arguably, the main difference between EU and pragmatic accounts of explanation may amount to a matter of emphasis concerning the possession of metacontent. EU involves not just knowledge of explanatory content and its employment, but also

[87] Similarly, Illari (2019) discusses a form of explanatory understanding, namely mechanistic understanding, and argues that: "A phenomenon is mechanistically understood when scientists have an intelligible mechanistic explanation for the phenomenon; i.e., a mechanistic explanation that *they can use*" (69; my emphasis).

possession of a higher order knowledge, namely knowledge about the metacontent and its management.

The modal status of genuine model explanations may affect the way they are used in explaining. For example, non-actual how-possibly model explanations are rarely taken to answer why-questions, and they are hardly effective in providing actual solutions to actual problems. Still, non-actual genuine model explanations can be used for other purposes. In particular, they can still be used as answers to questions of a different kind, namely *what-if-things-had-been-different questions* (cf. Woodward, 2003: 11). Various philosophers have related the ability to answer these questions to understanding. For example, Grimm (2017) argues that an important part of understanding is the ability to make modal inferences, i.e., the ability to identify how changes in one part of a system lead – or fail to lead – to changes in other parts (216-7). Ylikoski & Kuorikoski (2010) claim that the ability to answer what-if questions is a fundamental criterion to attribute understanding (205). And Reutlinger *et al.* attribute a modal function to how-possibly explanations which afford how-possibly understanding (1094).

This brings me to another downstream ability which is highly entrenched with the ability to answer what-if questions, namely *prediction*. Prediction is often an ability of EU, i.e., it relates to possession of model explanations. For example, in the case of deductive-nomological explanations, prediction is a natural extension of explanation. In fact, Hempel (1965) argues that deductive-nomological explanations provide understanding of why a phenomenon occurs in the sense of showing that it was to be expected (337). However, in other kinds of explanation, the explanandum is not entailed by the explanans and, hence, it is not predictable in the same sense as in the deductive-nomological case. Still, other kinds of explanatory knowledge can somehow be implemented in predictive endeavors. After all, explanations convey information about dependence relations which contribute to the overall predictive capabilities of a model. For example, knowledge of causal explanations, and in particular mechanistic explanations, provides content that is fruitfully implemented in predictive ventures, such as computer simulations.

In spite of the aforementioned relations between prediction and explanation, prediction does not necessitate explanation. Even if prediction relies on explanatory content, such content does not need to be explicitly known. These are cases in which models are used for predictive purposes – say, in simulations – but the agent lacks a genuine explanation of the results. In this case, prediction is not an ability of EU but rather an ability of OU. That is, prediction may be an ability by which agents relate to models without possession of model explanations. As an illustration of this circumstance, consider the case in which scientists use computer simulations to make predictions, but the model itself is a sort of "black-box" to them. In this case, the model is an effective tool, but it is not possessed by the agent ("possessed" in its technical sense).

The uses of models as simulations go beyond predictive ventures. A common kind of endeavor in simulations is that in which *type-knowledge*, gained by means of the content of the model, can be used to gain *token-knowledge* about a specific situation. This knowledge may, or may not, be explanatory knowledge. In this sense, this endeavor may indicate OU or EU. For example, consider a case discussed by Illari (2019) in which she explores the role of mechanistic model explanations of supernovae in simulations. Illari claims that explanatory knowledge about type-mechanisms of supernovae can be implemented in simulations. These simulations can be later put to use to learn about token supernovae. This endeavor demands adequate parameterizations and modifications to the simulation in order to better simulate the particular token. In the described case, we are entitled to attribute EU: a model explanation is possessed – the type mechanism – and it is employed for simulations of token mechanisms.

Often, models and model explanations are used *heuristically* (cf. Reutlinger *et al.* 2018: 1094; Frigg, 2003: 629). Roughly, this means that models and model explanations may provide suboptimal solutions to problems, which are considered acceptable given the constraints of the problems and the practical limitations. These solutions are often taken to be perfectible. In this sense, these successes can be conceived as steps forward towards an ultimate goal. Often, how-possibly and how-plausibly model explanations are used in this sense. They provide explanatory knowledge that is valuable but expected to be succeeded.

Finally, there is a family of uses of models and model explanations that refer to their broader and long-term impact in scientific inquiry. This idea has been variously described as the use of models in *developing new ideas* (Frigg, 2003: 629), as *starting points* (Gelfert, 2016: 84), for *developing better science* (de Regt & Gijsbers, 2017: 57), and the like. Models and model explanations may be used in various specific ways. But collectively, these specific uses amount to the broader goal of engaging in successful and enlightening scientific inquiry.

7.4 Précis

The main theoretical theses of this chapter are:

- Thesis 1 (EU): An agent has explanatory understanding of a target phenomenon with a model if and only if the agent possesses the ability – and its object – of crafting and/or using a genuine model explanation of the phenomenon, regardless of its modal qualification.
- Thesis 2 (OU): An agent has objectual understanding of a target phenomenon with a model if and only if the agent possesses the ability – and/or its object – of crafting and/or using an effective representational model of the target.

***

8 Understanding Differently and Its Exploratory Role

My account of scientific understanding with models, presented and discussed in the previous chapter, leaves various issues open for discussion. Still, I consider the account explicit enough for it to provide criteria for attributing understanding to scientists. In addition, it offers criteria for evaluating the attributed understanding as explanatory (EU) or objectual (OU), depending on scientists' possession of distinctive component elements of understanding (abilities and/or their objects). At the same time, it has to be admitted that the account is less explicit about criteria for appraising the attributed understanding in question. The problem is this. As hinted at in the last chapter, "understanding" is a gradual concept: scientists may possess more or less understanding about a certain target phenomenon. But then it is crucial to decide how one compares different states of understanding. My account does not provide sufficiently explicit criteria for resolving these issues.

In order to give a taste of the problems at hand, consider what follows from the attribution of understanding. Roughly, my account submits that understanding can be attributed to scientists in possession of abilities and/or the objects of these abilities. When the matter of attribution is resolved, one might ask: Does a scientist who possesses both abilities and their objects have more understanding than one who only possesses an ability or only its object? What about scientists who possess upstream abilities versus those who possess downstream abilities? What about those scientists who possess a specific set of abilities versus scientists who possess a different set? What about scientists who possess objects of abilities with different contents? Or with more or less active content? And what about those scientists whose explanatory understanding is based on how-actually explanations versus those whose explanatory understanding is based on how-possibly explanations? These questions address only some of the various differences that may obtain between states of understanding. Answers to these questions are needed in order to sustain the view that understanding is gradual. Unfortunately, I do not have all the answers to these questions. And I fear that any attempt to respond might inevitably end up in solutions of limited scope because of the contextual character of understanding.

This situation is less than ideal. Surely, philosophers of science are not only interested in the binary issue of whether scientists understand or do not understand a certain target phenomenon. Nor are they simply interested in evaluating such understanding as explanatory or objectual. Philosophers of science are also interested in deciding what makes the state of understanding of a particular (group of) scientists distinct from the state of understanding of other scientists. And if the states of understanding are distinct in these ways, which state of understanding is more valuable, even if only judged by local desiderata. I do not intend to solve these issues. The reasons for this decision are mainly practical: the evaluative comparison between states of understanding is an extremely complex task. It involves multidimensional analyses of various elements and contextual evaluations of the results of such analyses. This project is certainly worth pursuing, but its completion demands efforts that go beyond what I have been able to achieve in this dissertation.

Having conveyed the complexity of the problem, I make a first step towards dealing with it. In this chapter, I assess and compare states of understanding. Whether the results of this exercise serve as a model to be followed is left as a matter of further research. I set two main constraints. First, I conduct these assessments and comparisons in a specific context, namely the context of my case studies: the BK and OFC cases. Second, I focus on one dimension of analysis to compare BK and OFC's states of understanding. This dimension is a certain aspect of the content of their understanding, namely the non-representational content of their models. Third, I discuss an important aspect of having distinct states of understanding in terms of content, namely exploratory advantages.

The structure of the chapter is as follows. In the first part, I assess BK and OFC's possession of the various components of scientific understanding. This allows me to decide whether they possess understanding and to evaluate it as OU and/or EU. In the second part, I assess the differences between BK and OFC's states of understanding, focusing on their different contents. An important conclusion is that BK and OFC's states of understanding are significantly different. This is not a trivial conclusion, especially considering that the research in both cases is directed at the same target – earthquake occurrence – and is driven by the same underlying metaphor, namely seismic-faults-as-spring-block-systems. This circumstance

is described as an instance of understanding differently. In the third part, I suggest that the differences in content can be traced to differences in explanatory commitments that guide the construction and interpretation of the BK and OFC models. I suggest that these differences in explanatory commitments correspond to exploratory motivations. I provide a scheme for this thesis, which distinguishes between programmatic and prospective explorations. Without further ado, I proceed with the assessment and evaluation of understanding in the BK and OFC cases.

8.1 Assessment and Evaluation of Scientific Understanding in the BK and OFC cases

As I argued in the previous chapter, my account of scientific understanding is intended to be an externalist one. This implies that the assessment and eventual attribution of understanding to an agent are based on publicly accessible criteria or, at the very least, criteria that are susceptible to being publicly accessed (e.g., by interrogating the agent). This is, in fact, why I require explicit knowledge in the possession of the object of ability. Nonetheless, the fact that the criteria are openly accessible to the analysts does not mean that they will agree on their assessments. In other words, the assessment and attribution of understanding is highly contextual. Accordingly, disagreements on these matters are likely.

Having admitted the potential controversies surrounding this enterprise, I nevertheless engage in an assessment of the state of understanding in both the BK and OFC cases. In order to minimize arbitrariness, I attempt to restrict my assessment to material that is provided in the texts and figures of the relevant papers. In addition, I consider further references to these papers by other authors. In order to conduct this task, I proceed to identify and discuss the constitutive elements of scientific understanding with models, namely possession of: i) upstream abilities; ii) downstream abilities; and iii) objects of abilities. Once these elements are examined, I am in a position to evaluate understanding. I begin with the BK case.

8.1.1 The BK Case

With regards to upstream abilities, it is rather uncontroversial to attribute possession of them to BK. After all, they do construct a novel model of earthquakes, which is the definite expression of upstream ability. Furthermore, their model is not just another model within a well-established tradition but a fairly original one. Certainly, there are elements of the BK model that are taken from theories available at the time. But what BK do with these elements is outstanding. As BK themselves put it, their model embodies "a stupendous problem in general and one which has not yet been attempted" (342).

To better appreciate BK's upstream abilities, it is fruitful to analyze these abilities in terms of BK's understanding of theories. (As I argued in the previous chapter, theories are collections of principles which provide the basic tools for the construction of models. And the understanding of theories amounts to the ability of choosing and using theories to build models and model explanations, based on skills and judgments of the scientists in a given context.) I suggest that BK display a refined understanding of theories in the following sense. BK do not merely choose and use ready-made theories. Rather, they examine available theories and discriminate which of their constitutive principles are valuable for their project. This allows BK to assemble principles into a novel theory, which is used to construct an original model of earthquakes.[88]

There are two core principles underlying the construction of the BK model (see discussion in Chapter 3). The first principle is the suitability of a discrete system to represent and simulate the behavior of a continuum medium (P1). The second principle is the sufficiency of friction as a causal factor to account for the main features of the statistics of naturally

[88] In my account, a theory is a collection of principles that is used for model construction. Thus, a novel theory is a collection of principles that is novel in terms of introducing new principles, but also in assembling old principles from pre-existing theories in novel arrangements. Nonetheless, I admit that it is contentious to call any newly assembled collection of old principles a novel theory. Certainly, not all scientists would be willing to call an old theory with minor updates a novel theory. The question thus becomes how much change an old theory needs to undergo in order to be called a novel theory. I do not propose a definite criterion for this. In fact, a strict criterion may not even be desirable. Still, I suggest that, more than a matter of quantity, it is a matter of quality of those principles that are being changed and assembled in new arrangements. What I have in mind is similar to the criterion of identity of research programs, which is based on their core commitments (Chapter 3). Analogously, I think that a novel theory must be different from existing theories in terms of its core principles. Using this criterion, BK's theory is novel: they assemble a new arrangement of pre-existing principles and they play a central role in the construction of their model (see Chapter 3 for a discussion on the core-character of BK's assembled principles, P1 and P2).

occurring earthquakes (P2). Certainly, these two principles do not exhaust the theory underlying the construction of the BK model. In fact, other theoretical principles have been discussed elsewhere in this text. For example, various forms of idealizations (see Section 6.2.2) and techniques (e.g., Runge-Kutta method, referred to as P2 in chapter 3) are instances of principles that also shape the construction of the BK model. But the point is that these latter principles can be overwritten while keeping the core theoretical principles in place. That is, idealizations – other than the core ones – can progressively be overwritten by more realistic assumptions, and the various techniques used in the model could be replaced or perfected. What is at the core of the spring-block model is the conjunction of P1 and P2. I suggest this core is the key to appreciating BK's lasting contribution.

Given that upstream abilities may be based on tacit knowledge, it is relevant to recognize that BK's upstream abilities are not a good example of this. Quite the opposite: BK articulately express their reasons for assembling a new set of principles to study the occurrence of earthquakes. And state clearly what those principles are. In fact, BK spend an important part of their paper justifying their assumptions for the construction of the spring-block model. In this sense, BK's upstream abilities – i.e., their understanding of theory – are based on explicit knowledge of theory.

With regards to downstream abilities, the uses of the BK model span various achievements. To begin with a more strategical one, the BK model is used as a starting point for a new line of research. In spite of its idealizations – or actually because of them – the BK model is used to start reasoning about the occurrence of earthquakes in the form of a tractable mathematical problem. The fact that the BK model is a starting point in a progressively more refined line of research is evidenced by the amount of publications that refer to it, including OFC's work. Furthermore, words of praise have showered BK's work and their model. For example: "When Burridge and Knopoff (1967) proposed this model, it was the first mathematical treatment of earthquake rupture and a very important development. Since then, hundreds of papers have been published using this model or its variants" (Kagan, 2014: 16).

BK also display more specific downstream abilities which can be broadly conceived as different forms of surrogative reasoning. This is expressed in BK's treatment of their model. BK study their model by means of detailed examination and controlled manipulations in both the laboratory and computer implementations. By inspecting the relations between manipulations and simulation outputs, BK are able to infer relations between features of their model and the kind of phenomena that it produces. The lessons learnt in the context of the model are then used to reason about the occurrence of earthquakes. Furthermore, these lessons are – at times – expressed as lessons about the occurrence of earthquakes, even if with reservations. For example, consider the following quote: "Our models have shown that the stress distribution on one fault is readjusted by the occurrence of an earthquake event on the same or on an adjacent fault" (371). In this statement, BK argue that they have learnt something about stress distribution on a fault by means of studying the model. This is a textbook example of surrogative reasoning.

A particularly important form of surrogative reasoning in the BK case is model explanation. BK attempt to account for aspects of earthquakes occurrence by means of lessons learnt in the context of their model. Certainly, given the highly idealized nature of the BK model, these explanations cannot be expected to be how-actually explanations. But BK know this and express it in various places in the paper (e.g., "[t]he extension of these elementary observations to the interpretation of the seismic dynamics of any real region is probably not justifiable at present": 370).

Still, BK take their model to provide explanations. According to my account of model explanations, these explanations can be taken to be genuine if they are adequately qualified with a modal term. I argue that this is exactly what BK do. For example, consider the following quote: "[I]f the demonstrations of the laboratory and numerical models are borne out in nature, it *would seem likely* [i.e., plausible] that the nature of the friction on a fault surface *determines* the statistical properties of the earthquake shocks that are observed" (370; my emphasis). In this case, the "would seem likely" is the literal modal qualification to the proposed explanatory relation of "determination" between friction and statistical features of earthquakes. Or consider this other claim: "The uniformity of values of b in [the Gutenberg-Richter law] observed over widely different regions of the world, *give rise to the possibility*

that, at least for large shocks, the frictional properties to be found on faults are relatively universal" (370). In this case, BK insinuate that relatively universal frictional features on faults how-possibly explain the uniformity of b values. So qualified, I submit that this a genuine explanation derived from the BK model.[89]

The objects of abilities in the BK case are mainly of two sorts. On the one hand, the object of most of the aforementioned abilities is a representational model, namely the BK model. Indeed, BK design and construct the model (i.e., object of upstream abilities) and use it in various forms as a surrogate system for studying the occurrence of earthquakes (i.e., object of downstream abilities). On the other hand, a set of more specific objects of abilities are those model explanations derived from the BK model. BK craft specific model explanations of different aspects of the occurrence of earthquakes, from their power-law distribution to the statistics of aftershocks. And BK are able to use these model explanations not only for purposes of explaining, but also for further inferential and predictive efforts.

To evaluate possession of these objects, one must assess BK's relation to their content and metacontent. Recall that possession of the object of ability amounts to: i) explicit knowledge of the active content; ii) explicit knowledge of the metacontent; and iii) knowing how to assemble the active content and adjust its metacontent in a way that secures coherence. Thus, I proceed to evaluate the obtaining of these conditions.

First, BK display explicit knowledge of the content of their model and model explanations. This is evident even from a superficial reading of their paper. BK explicitly assert propositions that are true of their model or in accordance with it. These propositions span content about the model's structural features, underlying theoretical assumptions, simulation results, and its representational relations to the target. This explicit knowledge of content is not only evident in BK's prose. It is also attested in the graphs that present

[89] As a matter of fact, the latter is not a how-actually explanation. To begin with, the explanandum is nowadays discredited: b values vary widely across tectonic regimes (see e.g., Schorlemmer *et al.* 2005). In fact, OFC themselves explain the opposite in their paper, namely why the b values in the Gutenberg-Richter law vary across seismic faults and regions. Secondly, the explanans is non-actual: the frictional properties on faults are not relatively universal (e.g., see "weak faults" in Collettini *et al.* 2009).

simulation results or the diagrams that illustrate the architecture of their model (see Chapter 2).

Second, BK also display explicit knowledge of the metacontent of their model and model explanations. Indeed, BK do not merely present content as if it were a collection of bullet points. Rather, they assert nuanced relations between the different portions of the content and even of a counterfactual sort. This is expressed in the various well-structured arguments, analyses and interpretations of results, and qualifications of their claims. BK's explicit knowledge of metacontent is particularly evident as one reads their explications of their model explanations. BK express how aspects of the model relate to each other and how their model can be used to develop explanations of features of earthquake occurrence based on the simulation results.

Thirdly, BK are responsible for assembling the active content of their model and model explanations and managing its metacontent. This is obviously the case since their model is a novel contribution: no one else has done it for them. But beyond this inference, BK actually explicate their reasoning in selecting the active content and managing the metacontent. As an example of the reasoning in selecting active content, consider the following passage: "A number of statistical properties of this system [i.e., the laboratory implementation the BK model] can be recorded. One of the principal features of interest is the potential energy in the system released during any shock" (344). This quote expresses BK's decision to focus on one aspect of their model among many. Later in their paper, BK explain why this content – the model's potential energy at different stages of the simulation – is worth being taken as active content: "We can also investigate frequency relations in the quake series. The logarithmic number-magnitude relation for earthquake occurrence in a given area is familiar [the Gutenberg-Richer law]. We have no direct measure of shock magnitude in our experiments. But we are able, in fact, to bypass the question of shock magnitude and deal directly with the potential energy released in the shock" (346). That is, BK use difference in potential energy between different stages of the simulations as a proxy for shock magnitude. This allows BK to use their model and its content to reason about an empirical-statistical law concerning earthquake occurrence.

In light of the preceding assessment, the following evaluation of BK's state of understanding seems natural. BK have objectual understanding (OU) of earthquake occurrence via the spring-block model. I attribute OU to BK given that: i) BK craft an original model (i.e., possess upstream abilities); ii) BK are able to use it in various forms (i.e., possess downstream abilities); and iii) BK assemble the content of their model, manage its metacontent and explicitly know both of them (i.e., possess the object of ability). In addition, BK also have explanatory understanding of certain aspects of earthquake occurrence via the spring-block model. In particular, BK gain explanatory understanding about why the number-magnitude distribution of earthquakes follows a power-law. This is the case because: i) BK craft genuine model explanations based on the spring-block model, in the form of a how-possibly mechanism; ii) BK explicitly know the content and metacontent of these explanations, as it is manifest in the lengthy and detailed descriptions of this mechanism; and iii) BK use these model explanations, whether for actual explaining, proposing applications for earthquake prediction, or sketching future developments.

8.1.2 The OFC Case

An analogous analysis can be conducted in the OFC case. With regards to their upstream abilities, I suggest that OFC possess them, given that they also construct a new model of earthquakes. Nonetheless, it is worth remembering that the OFC model is a modified version of the BK model. As OFC themselves acknowledge: "This model [i.e., the OFC model] is equivalent to a quasistatic two-dimensional version of the Burridge-Knopoff spring-block model of earthquakes" (1244). In this sense, it can be argued that OFC's upstream abilities are built upon BK's upstream abilities. Or, to put it differently, OFC adopt an important part of BK's theoretical understanding, at least with regards to their core theoretical principles. This does not diminish OFC's possession of upstream abilities. After all, OFC add new assumptions to BK's theoretical understanding and even overwrite some of its aspects. Because of this, I suggest that the OFC model can be regarded as a new model, in continuity with BK's core theoretical understanding, but with clear departures reflecting upstream ability.

Nevertheless, the novelty of the OFC model can be – and has been – put into question. There are two senses in which this case could be made. Firstly, one could claim that there are no significant differences between the OFC model and the BK model. For example, Pruessner (2012) claims that "[t]he [BK] Model remains the mechanical equivalent of the OFC Model, so *in that sense* the OFC Model is not a new model" (126; my emphasis). I highlight the words "in that sense" because they provide the right qualification to the claim of lack of novelty. The OFC model is not a new model insofar as it adopts BK's core theoretical principles and capitalizes on the same metaphor of the "spring-block" system. This covers the "mechanical equivalence" to which Pruessner refers. Nevertheless, I suggest that the OFC model is different from the BK model in at least two senses. First, it is based on a different vehicle, namely a cellular automaton computer simulation. This is quite unlike the vehicles in the BK model, namely: i) the material implementation of the spring-block system in the laboratory, and ii) the computer simulation which implements a numerical technique to solve the equations of motion of the spring-block system. Second, the OFC model affords a new interpretation of the spring-block metaphor. While BK's interpretation is mechanistic, OFC's interpretation is more abstract, recasting the problem as a mathematical one. I discuss this issue in more detail below.

Second, even if the OFC and BK models are significantly different, a more nuanced case could be made about the relative lack of novelty in the OFC model. From a historical perspective, the trajectory from the BK model to the OFC model involves several intermediate models, including other ventures by the same authors. For more details on this historical reconstruction, see Pruessner (2012: 125-9). The relevant point is that various of the modifications and upgrades that OFC introduce to their version of the BK model had already been presented in other models. In this sense, the credit – and claims of novelty – may need to go to these intermediate links. The OFC model capitalizes on various contributions that engaged with BK's work. In fact, OFC acknowledge these efforts (Christensen & Olami, 1992b: 8729). But this does not diminish the novelty of the OFC model. There is a sense in which novelty is attained in terms of synthesis and formality. Pruessner (2012) remarks on this: "Inspired and guided by this work [i.e., the various responses to the BK model] Olami, Feder, and Christensen (1992) finally generalized the [BK] Model comprehensively and formulated it most succinctly along the lines of a continuous sandpile model" (126). Furthermore, there is

at least one major theoretical upgrade that OFC introduced, namely the identification of an explicit parameter that regulates the conservation level (see α in Chapter 2). By regulating this parameter – and thus setting their model into conservative and non-conservative regimes – OFC conclude that the SOC behavior of their model is non-universal. Thus, I reaffirm my claim that OFC display upstream abilities. They are not just borrowing an already existing model. They are constructing a new model based on previous efforts but with clear departures that manifest theoretical understanding.

As with BK, OFC's upstream abilities are also based on explicit knowledge of the theory guiding the construction of their model. This is evident, for example, in OFC's explicit presentation of the algorithm that their model embodies, namely the instructions for the driving and relaxation stages (1245). But the clearest illustration of OFC's explicit awareness of their theory is embodied in their comparisons to other related theories, particularly two. One is BK's theory from which OFC explicitly depart in various regards. Particularly important are the two-dimensional extrapolation of the OFC model and its quasi-stationary procedure (see "separation of timescale" in Chapter 2). The other theory to which OFC respond is the overall theoretical framework of the SOC program. Particularly important in this regard are two principles that challenge the received theoretical views in the SOC program, namely the assumptions of continuous state variables and non-conservative relaxation.

In terms of downstream abilities, OFC use their model for various purposes. To begin with more general and strategic uses, the OFC model is employed to open new lines of research. This is the case in at least two senses. First, the OFC model points at new problems to be researched in the context of the SOC program. This is the case given that the OFC model simulates SOC behavior by assuming theoretical principles that were once considered outside of the scope of SOC. Second, the OFC model also prompts lines of research for professional seismologists, who are compelled to respond to the SOC-inspired upscaling techniques deployed in the OFC model.

The OFC model is also used for more specific purposes. In analogy to my analysis of the BK case, most of these specific uses can be collectively characterized as different forms of surrogative reasoning. Roughly, surrogative reasoning amounts to studying the model,

learning about its dynamics and then imputing some of these lessons to the target, in this case seismic faults. OFC clearly express these surrogative motivations for their model. After all, the OFC model is presented as a *model of earthquakes* and it is used to gain insights about earthquake occurrence. OFC are justified in engaging in surrogative reasoning given that their model "exhibits the features of real earthquakes" (Christensen & Olami, 1992b: 8729; see "phenomenal resemblance" in Chapter 6). They claim elsewhere that they have "good reason to believe that this simplistic picture [i.e., the OFC model] has a real connection to the actual fault dynamics that leads to earthquakes" (Christensen & Olami, 1992a: 1835).

Given the various idealizations embedded in the model (see Chapter 6), the imputation of lessons learnt in the context of the model to the target need to be adequately qualified. In this regard, OFC display laudable lucidity as they reflect upon the surrogative character of their research and its scope: "In the realm of experimental physics, physicists deal with the characteristic behavior of real physical systems. In the world where theoretical physics reigns, physicists are concerned with simplified models. The majority of these models are, of course, derived from real physical systems. The golden rule, when mapping a physical system into a model system, is to grasp only the important features of the relevant phenomena. Otherwise, the model system can very easily turn out to be too complex, so that it will be almost impossible to comprehend the mechanism, which is responsible for the observed behavior" (Christensen & Olami, 1992b: 8729).

Some of the outcomes of surrogative reasoning are used to sustain claims of potential predictive capabilities. For example, consider the following argument. Christensen and Olami (1992b) notice that small earthquakes (in the simulations) do not correlate in time. They argue that this is because small earthquakes are not correlated in space: a small event in one part of a seismic fault does not influence the occurrence of a small event in another part of the fault. However, if one studies subsections of the fault, then correlations between these smaller events appear. Christensen and Olami conjecture that "it might be very useful to pay attention to the much more numerous small events to get some statistical *predictions* for the few large events!" (Christensen & Olami, 1992b: 8733; my emphasis). And in their conclusion, they state that: "We may be able to improve our ability to estimate the probabilities of occurrence of large earthquakes, even with the limited information available about the

history of earthquakes occurrence and existing faults by utilizing the information hidden in the numerous small events!" (1992b: 8734). It is worth noticing the qualifications of these potential predictive capabilities ("it *might* be very useful" and "We *may* be able"). OFC also claim that their model "gives a good *prediction* of the Gutenberg-Richter law" (1244; my emphasis). However, in this context, "prediction" seems to be synonymous with "simulation". That is, the OFC model predicts the Gutenberg-Richter law in the sense of being able to simulate it.

An important form of surrogative reasoning in the OFC case comes in the form of model explanations. In fact, OFC are more outspoken about the explanatory purposes of their work than BK. OFC explicitly say that their model explains: i) the observed power-laws in earthquake occurrence; and ii) the variances in the observed b values across seismic faults and regions (OFC: 1244; Christensen & Olami, 1992a: 1835). Focusing on the later explanandum, variances in the observed b values are explained by variances in conservation regime (see Fig. 2.12 in Chapter 2). OFC do not qualify this explanation in their original paper, at least not explicitly. However, in their subsidiary papers, they do provide some hedging. For example, Christensen & Olami (1992a) say that their results give "some explanation" to the observed variability of b values (1835). And in Christensen & Olami (1992b), they claim that "the observed variation in the b value in the Gutenberg-Richter law *could be explained* by a variation in the elastic parameters" (8729; my emphasis). And again "[t]his variability [i.e., changes in α] *could serve as an explanation* for the variances in the observed B" (8731; my emphasis). These qualifications suggest that the model explanation of variances of b values is not intended to be a how-actually explanation. Most likely, it is intended as a how-possibly explanation, given the stark idealizations embedded in the model (of which OFC are fully aware; see above).

In parallel with the BK case, the objects of abilities in the OFC case are two. One object is the OFC model itself. Indeed, the OFC model is crafted by OFC, making it the object of upstream abilities. And the OFC model is used by OFC in various forms, which makes it the object of downstream abilities. The other object of abilities is the set of model explanations that OFC derived from their model. Model explanations are the product of upstream abilities and are utilized in downstream abilities. Particularly salient model explanations are the

explanation of the Gutenberg-Richter law and the explanation of the variances of b values in the Gutenberg-Richter law.

I argue that these objects are possessed by OFC. First, OFC explicitly know (part of) the content of their model and model explanations. This explicit knowledge is reported in OFC's papers. Consider the detailed presentation of the OFC model, which includes the underlying theoretical assumptions, structural features, simulation results, and interpretation of them. Explicit knowledge of the content of the OFC model is not only presented in the text: OFC attach several diagrams and plots that allow them to report the structure of the model and the results of simulations. The content of model explanations is also reported explicitly. As argued above, OFC voice some of the results of their model as explanations – even if how-possibly explanations – of target phenomena. This is clearly the case with the explanation of the variance of b values, which are reportedly explained by variances in conservation regimes, controlled by the α parameter.

Note, however, that there is a model explanation presented as possessed, but not explicitly reported: the explanation of the power-law behavior. Consider the following quote: "Thus our results, apart from providing an explanation for the observed power laws, also give some explanation for the observed variability [of b values]" (Christensen & Olami, 1992a: 1835). I argued above that the variability of b values is explicitly explained by resorting to resources in the OFC model, namely the α parameter. Still, an explanation of the power-laws is absent, at least in the form of an explicit formulation. I surmise that OFC assume that an explanation is being given by the very fact of constructing and describing a model that simulates the power-laws. However, as I argued in Chapter 7, merely pointing at a model with the right explanatory content does not amount to explaining. Unfortunately, OFC do not make explicit what features of their model are responsible for bringing about the power-law behavior.

I conjecture that OFC are implicitly assuming the received explanatory strategies of the SOC program for the power laws but are not really spelling them out. There are explicit proposals of such explanations in the literature. One example that might have influenced OFC's views states: "The explanation is that open, extended, dissipative dynamical systems

may go automatically to the critical state [i.e., power-law behavior] as long as they are driven slowly: the critical state is self-organized" (Bak & Chen, 1989: 5). Or consider the more recent "SOC's genotype" approach (cf. Watkins *et al.* 2016: 21-2; see also discussion in Chapter 2). Certainly, it is likely that OFC explicitly know the content of explanations of power-law behavior, as they display expert knowledge of the SOC literature. But the point is that OFC do not report these explanations explicitly. Given this lack of evidence, it remains controversial to attribute possession of this particular model explanation to OFC.

Second, OFC explicitly know the metacontent of their model and model explanations (at least the explanation of variance in b values). This is certainly evident in the various elaborate arguments that they construct and the discussions in which they engage. But a more conspicuous piece of evidence comes in the form of cross references. Recall that, in the same year, the OFC model was discussed in three papers by their authors: i) Olami, Feder, and Christensen (1992); ii) Christensen & Olami (1992a); and iii) Christensen & Olami (1992b). Each of these papers addresses specific aspects of the model's content. By cross referencing, OFC point at the interrelations between portions of content, thus displaying explicit knowledge of the metacontent of their model and model explanations.

Third, OFC are responsible for assembling the active content of their model and model explanations and managing its metacontent. (The case is mostly analogous to that of the BK case.) Conveniently, OFC explicate how they engage in these tasks. With regards to assembling the active content, OFC make it clear that the main criterion is the relevance of the content to the SOC program and the study of earthquake occurrence. With regards to managing the metacontent, OFC display this capability as they adjust and qualify their claims as responses to the existing literature in SOC and earthquake research. In fact, responding to the earlier literature by offering a new model may well be the epitome of management of metacontent.

Thus, in light of the previous assessment, I am in a position to evaluate OFC's state of understanding. OFC have OU of earthquake occurrence via their model. This attribution is based on three considerations: i) OFC craft a new model (i.e., possess upstream abilities); ii) OFC use this model for various purposes (i.e., possess downstream abilities); and iii) OFC

assemble the content of their model, manage its metacontent and explicitly know both of them (i.e., possess the object of ability). In addition, OFC have EU of certain aspects of earthquake occurrence via their model. It is clear that OFC have explanatory understanding of the variances in b values of the Gutenberg-Richter law across faults and regions. This is because: i) OFC craft a genuine how-possibly model explanation; ii) OFC explicitly know the content and metacontent of this explanation; and iii) OFC use the model explanation. A less convincing case could be made about OFC's explanatory understanding of power-law behavior, given that they do not explicitly report the explanation. Still, they construct a model that simulates the power-law behavior and they explicitly know its inner structure. In that sense, one could surmise that OFC may also have EU of power-law behavior of earthquakes.

8.2 Understanding Differently

In both cases, BK and OFC acquire objectual and explanatory understanding of earthquake occurrence by means of their respective models. But this does not mean that their understanding of earthquake occurrence is the same. Certainly, their understandings are of the same type. But their understandings are attributed based on different components. In other words, scientists may have the same type of understanding, even towards a same target phenomenon, without having the same *state of understanding*. The state of understanding amounts to the particular constitution of the understanding attributed to an agent, whether EU or OU. It relies on the component elements of the attributed understanding, namely the abilities of understanding and the objects of such abilities. Dissimilar abilities and objects – even starkly different – may enable the attribution of the same type of understanding towards a same target phenomenon to different scientists who possess them.

I argue that this is exactly the case with BK and OFC. Their states of understandings towards a same target phenomenon are of the same type, namely EU and OU. But, a fine-grained examination of the components of these states reveal significant differences between their states. Some of these differences have been pointed out above and in previous chapters. It is not my intention here to provide an exhaustive review of all the differences between BK and OFC's states of understanding. Instead of engaging in such multidimensional assessment,

I focus on mainly one component, namely the objects of abilities, that is, the models in question and model explanations.

Certainly, there are several similarities between the BK and OFC models. Both are presented as models of earthquakes which simulate aspects of their occurrence, particularly their power-law behavior. These models are also similar in terms of their structure. In fact, the OFC model is introduced as a version of the BK model. This implies that OFC adopt core principles of the BK model to construct their own model. And – arguably the most striking similarity – both models rely on the same general metaphor, namely that which portrays seismic faults as spring-block systems. In fact, OFC present their model as "*equivalent* to a quasistatic two-dimensional version of the Burridge-Knopoff spring-block model of earthquakes" (1244; my emphasis).

In spite of these similarities and connections, the BK and OFC models are ultimately different in significant regards. A revision of my discussion above or in previous chapters should suffice to convince the reader. My goal in this section is to spell out a specific sense in which these models are different, namely in terms of their *type of non-representational content*. The non-representational content of a model is the set of propositions that are true about the model in accordance with a non-representational interpretation of its vehicle. Given that the non-representational interpretation of vehicles is guided by explanatory commitments (see Chapter 5), I suggest that the type of non-representational content of a model is that of the explanatory commitments that guide the interpretation of its vehicle.

At this point, it is relevant to recall my argument about explanatory commitments and accounts of explanation. In Chapter 4, I construed accounts of explanation as stable sets of explanatory commitments. I noted two problems in using accounts of explanations to describe the explanatory practices of scientists. First, scientists' practices may embody explanatory commitments at the intersection of distinct accounts of explanation (the problem of overlaps). Second, their practices may embody explanatory commitments taken from different accounts of explanation (the problem of integration). Because of this, using accounts

of explanation to describe explanatory practices is a suboptimal solution.[90] However, it is a solution that is widely used among philosophers of science to convey information about explanations.

Now, consider the issue of type of content. There is an apparent problem in saying that the type of non-representational content of a model is that of the explanatory commitments that guide the interpretation of its vehicle. That is, the explanatory commitments need not be – in any straightforward sense – of a particular type: there may be conflicts in allocating the explanatory commitments to specific accounts of explanation. Instead of a problem, I take this consideration to provide constraints on what is meant by "type of content". The notion of "type of content" mimics that of "account of explanation" in being a stable category for the purpose of description. Thus, describing content in terms of its type inherits the problems of overlaps and integration. Still, the notion of type of content proves useful – as I show below – in pointing at significant differences in content. With these caveats in mind, I proceed with my argument.

I have argued above that BK and OFC have different explanatory commitments. In Chapters 4 and 5, I suggest that BK and OFC's explanatory commitments can be characterized as mechanistic and mathematical, respectively (with all the nuances that these claims deserve). These commitments are reflected in BK and OFC modeling decisions. In other words, the BK and OFC models embody mechanistic and mathematical non-representational interpretations of their respective vehicles. To be clear, this does not mean that the BK and OFC models may not be re-interpreted in ways that depart from the original mechanistic and mathematical interpretations of their authors. It only means that these are the interpretations that their authors express in their treatments. As a consequence, the types of

[90] Rice *et al.* (2018) submit a similar thesis. They suggest that focusing on the description of explanatory practices with models in terms of well-established philosophical accounts of explanation misses something important. They argue that besides this enterprise, philosophers of science should also focus on describing "explanatory schemas," which transcend specific accounts of explanation. However, I think one problem with their proposal is that they do not consider that part of the justifications that scientists use are practical justifications, based on explanatory commitments. That is, scientists justify their modeling practices based on how they aim to engage in explanations with such model. To be fair, this seems like an "egg-chicken" problem. In that sense, I think that insisting in the problem misses the point. This is why, in the end, I still resort to accounts of explanation to describe scientists' practices, even though I am critical of how that is done.

non-representational content of the BK and OFC models are mechanistic and mathematical, respectively.

To appreciate the significance of different types of content in these models, consider BK and OFC's treatment of the common metaphor underlying their models, namely that of "seismic-faults-as-spring-block-systems". BK introduce the idea of representing seismic faults as spring-block systems mainly due to two reasons. First, the spring-block system is minimal: it isolates friction as a causal factor. Second, it is tractable: insofar as it is discrete, it enables controlled observation, experimentation, and mathematical treatment. These simplifications enable BK to study the occurrence of earthquakes.

Within the context of these simplifications, BK's treatment of the problem remains physical, in particular mechanistic. This is manifest in the material implementation of the spring-block system in the laboratory, where an actual spring-block mechanism is constructed and experimented with. But it is also manifest in the computer implementation, in which the computer program describes the spring-block mechanism in terms of the equations of motions of its blocks. And it simulates the behavior of the spring-block mechanism by solving (numerically) the system of equations that describe the mechanism. To be sure, mathematics is certainly used in the BK model, not only to describe the equations of motion, but also to solve them by means of a numerical method. However, the point is that the mathematics is not carrying the explanatory force of the model. It is the physical description of the spring-block mechanism which provides the BK model with its explanatory force.

OFC also use the spring-block metaphor, but they further idealize it, representing the spring-block mechanism as a purely mathematical entity, that is, as a cellular automaton. In this treatment, no physical theory is implemented. They describe this mechanism to the extent that it can be captured by coarse arithmetical instructions, leaving aside much of the underlying physics. Because of this, the OFC model can be characterized as a model of a model of earthquakes. The frictional thresholds in the BK model become a stipulated maximum for the value of variables in OFC's cellular automaton. And the relaxation, which is the result of solving physical equations in the BK model, becomes a stipulated arithmetical instruction in the OFC model. Frigg (2003a) makes a similar point, although his claims have a broader scope,

addressing SOC models in general. Frigg argues that SOC models, in the form of different cellular automata, embody a mathematical calculus. Certainly, these calculi can be thought of as representing physical systems (in the OFC case, a spring-block system). But the calculi are not themselves solving physical theory.

As a result, the non-representational content of these models is different. The non-representational content of the BK model is physico-mechanistic. That is, there are propositions that are true about the BK model that state physico-mechanistic facts about the model (not considering its representational role). For example, "after a considerable length of time […] large shocks occur in which all eight masses move" (BK: 345). Or, "the logarithm of the number of shocks with potential energy release greater than E against the logarithm of E […] is essentially a straight line with slope -1 over a wide range of energies" (347). These propositions are true of the BK model in one or more of its simulations.

In contrast, the non-representational content of the OFC model is mathematical: the OFC model is a cellular automaton. Thus, no propositions stating physico-mechanistic content are true of it. It is worth noticing that – nonetheless – OFC commonly resort to physico-mechanistic terms to describe their model. For example, OFC often talk about "blocks," "forces," and even "earthquakes" to describe what is happening in the simulations. My reading is that, in using these terms, OFC are describing their model in terms of its representational content, as opposed to its non-representational content. Indeed, OFC make it explicit that their intention is to map their cellular automaton to a two-dimensional version of the spring-block system and – through the latter – to seismic faults. This is an intuitive way of describing the OFC model, and it plays a communicational purpose. I insist: no propositions stating physico-mechanistic content can be true of a cellular automaton.[91]

[91] Recalling a reflection by Gunnar Pruessner in an interview (personal communication), the OFC model is a spring-block model in the same sense that the plum pudding model of the atom is a model of plum puddings. The model just resembles its label and this resemblance serves as a mnemotechnic device. The OFC model, in that sense, should not, strictly speaking, be considered a spring-block system, but a mathematical model that tests the robustness of outputs in an abstract calculus. There are several systems whose behaviors loosely resemble the procedure of such calculus, e.g., a spring-block system.

Thus, BK and OFC's states of understanding of earthquakes are different in terms of the non-representational content of their objects of abilities. This difference has a lasting impact in the overall content of BK and OFC's understanding of earthquake occurrence. Indeed, it is different to understand earthquakes from the perspective of a mechanistic or mathematical model. As shorthand, I say that BK have "mechanistic understanding" of earthquake occurrence and OFC "mathematical understanding" of it. Mechanistic understanding is understanding whose object of ability is a mechanistic model. Mathematical understanding is understanding whose object of ability is a mathematical model. Note that mechanistic and mathematical understanding may or may not involve explanations. This is because mechanistic and mathematical understanding may occur in explanatory but also purely objectual understanding.[92]

A central question for philosophers of science is why scientists engage in understanding differently a same target phenomenon, i.e., representing the target by means of models with different non-representational content. I suggest that the answer involves, at least partly, the following consideration. Possession of objects of ability – whether model or model explanation – with different non-representational contents allows the users to think about the target in different ways. In other words, the users may be able to engage in different forms of surrogative reasoning. In these tasks, the reasoning process is different because its raw material – i.e., the content – is different. And different reasoning processes may – but do not need to – lead to different conclusions. These conclusions may not all have the same epistemic status, but that is not my concern here. The point is that by understanding differently with models, scientists generate alternative ways to reason about their target. To put it in a few words, understanding differently plays an exploratory role. In the following section, I sketch a proposal for further research on the relations between understanding differently and scientific exploration.

8.3 Scientific Exploration as Attempts to Understand Differently

[92] This approach contrasts with that of Illari (2019), in which mechanistic understanding of a phenomenon requires having an intelligible mechanistic explanation for the phenomenon, i.e., a mechanistic explanation that can be used.

8.3.1 Notions of Exploration

Exploration is a comparatively neglected topic in philosophy of science. This is surprising given the crucial role that exploration plays in scientific inquiry. In particular, the role of models in scientific exploration has only recently been discussed (e.g., Gelfert, 2019; Massimi, 2018; Reutlinger *et al.*, 2018; Gelfert, 2016). Given this deficient background, I first clarify what I mean by exploration. Here, I consider Gelfert (2016) the most helpful: "[…] exploration is simply an activity that aims at the discovery of new facts, with the term 'exploration' designating a behavioral pattern and 'discovery' referring to a new epistemic advance in our cognitive state" (74). Acknowledging that discovery has received more systematic attention than exploration, Gelfert embarks on the investigation of the latter notion. And so do I.

For the purposes of my discussion, I focus on three related concepts that cover most of what I intend to say about exploration, namely those of "specific exploration," "isolation," and "indirect exploration." The notion of "specific exploration" I adopt from Gelfert (2016). Based on Berlyne (1960), Gelfert submits that there is "specific" and "diversive" exploration. On the one hand, specific exploration is driven by scientists' interest in a novel, unexpected, or not fully understood phenomenon. The explorative tasks in specific exploration can be characterized as convergent: they focus on a specific target. On the other hand, diversive exploration is shaped by the search for novel and unexpected stimuli. In other words, diversive exploration does not aim for the discovery of new facts about a specific target. Rather, diversive exploration aims for the discovery of new facts *simpliciter*. In diversive exploration, explorative tasks are valuable for their own sake.

Closely related to Gelfert's notion of "specific exploration" is the notion of "isolation", as presented by Mäki (2009a, 2009b). Both notions stress the importance of focused attention on specific variables, undisturbed by other factors. Mäki discusses isolation in the context of idealized scientific models. He argues that idealizations in models are falsehoods about their targets, but falsehoods that should not be taken as mistakes. Rather, idealizations are used strategically by scientists and serve a deliberate purpose, namely that of "theoretically *isolating* causally significant fragments of the complex reality" (Mäki, 2009a: 71; my

emphasis).[93,94] In the case of theoretical models, isolation amounts to neutralizing the effect of certain parameters via idealizations. In the case of material models, isolation involves the causal controlling of noise and non-focal causal factors that affect the obtaining of a target phenomenon and its particular features (see Mäki, 2009b: 30).

A third important notion is that of "indirect exploration". The basic idea is that target phenomena can be explored indirectly via the direct exploration of a model that represents it. Accordingly, indirect exploration is a form of surrogative reasoning. Mäki (2009b) explicates the motivations behind this notion. In Mäki's view, exploration is a scientific activity conducted on a model. This activity basically consists in the examination of the properties of a model and checking the behavioral implications of the model having such properties (Mäki, 2009b: 34). This way, inferences about the behavior of the model can be drawn. Eventually, exploration conducted on a model can be used to indirectly learn about a target system. A main reason to engage in indirect exploration is that, in several cases, it is a more expedient strategy than the alternative direct examination of the target phenomenon. This is certainly the case, if one considers that models are intended to isolate focal causal factors of interest to the scientists. Thus, there is a deep connection between indirect exploration and isolation.

Further relations between isolation, specific exploration, and indirect exploration can be examined in the context of my case studies. Consider the BK case. BK explicitly use the terms 'exploration' and 'isolation' to characterize their model-based study of earthquakes. Furthermore, their usage of these terms suggests that isolation relates to exploration as a means to its end. For example, BK claim that "[a] laboratory and a numerical model have been constructed to *explore* the role of friction along a fault as a factor in the earthquake mechanism" (BK: 341; my emphasis). Thus, exploration is the end. Then, BK declare how isolation relates to this end: "The feature we choose to *isolate* is the conjecture that a gross form of friction between the two walls of an earthquake fault inhibits the relative

[93] Mäki develops most of his philosophical ideas about modeling in science in the context of economy. However, several of his insights can be safely, and even fruitfully, extrapolated to other scientific disciplines.
[94] Mäki often refers to these "causally significant fragments" as "mechanisms" (e.g., Mäki, 2009a: 78). Mäki does not declare alliance with the new mechanists in his employment of the term 'mechanism'. However, I submit that the minimal account of mechanism that I introduced above (Chapter 4) is compatible with Mäki's notion of "mechanism".

displacement of the material on the two sides" (BK: 342; my emphasis).[95] From the previous quotes, it follows that BK express clear exploratory motivations and isolation plays a role in achieving these aims.

BK's exploration can be described as specific and indirect. They engage in specific exploration insofar as BK explore a specific target phenomenon, namely the role of friction in the production of earthquakes in seismic faults. This exploration is indirect insofar as BK explore this target phenomenon via the direct exploration of the spring-block model (an instance of surrogative reasoning). Indeed, BK explicitly say that they have constructed their spring-block model to explore the role of friction (341). The merits of exploring the model rather than directly exploring the target are straightforward: the BK model isolates friction as a causal factor. This allows for a focused study of the role of friction, undistracted by other factors.[96,97]

Still, BK's exploratory efforts go beyond merely studying the role of friction by isolating it in a model. One has to recall that, at the moment when they were dealing with the problem of earthquake occurrence, no physical theory was available to them (at least not in analytical form). It is one thing to design models when one works within an accepted theory or even a working theory. Another thing is to model with no theory available. Indeed, BK had to put together some principles in a novel way to build their model. This is a case in which modeling is exploratory in a deeper sense, or so I suggest in the next section.[98]

---

[95] In addition, consider these other quotes mentioning "isolation": "In this paper we wish to *isolate* one of the qualitative features used in one of the models of earthquake occurrence [namely, friction]" (BK: 342; my emphasis). "We wish to consider the problem of friction along an earthquake fault as an *isolated* problem" (BK: 342; my emphasis).

[96] Isolation of friction is not the sole reason to prefer exploration of the BK model over the direct exploration of the target (i.e., earthquakes in geological faults). There are three additional considerations that make exploration of the BK model preferable over the direct exploration of its target. First, there are problems of accessibility. In particular, most of the extension of earthquake faults are subterraneous. Second, there are problems of manipulability. Frictional forces, differential tectonic stresses and elasticity of rocks cannot be (easily) intervened in real earthquake faults as in the BK model. Third, the kinematic of real earthquakes is slower than that of laboratory and numerical models. In this sense, scientists can get more results in less time by running simulations than they would obtain by observing real earthquakes.

[97] To be sure, friction is isolated *along* with other causal factors, such as the differential tectonic stresses represented by the pulling of the motor and the elasticity of rocks represented by the springs connecting blocks. BK isolate friction, but this does not mean that they consider friction solely. Rather, it means that friction is taken as the *focal* causal factor in the model.

[98] Gelfert (2019), in discussing exploratory experimentation, argues that "often enough – especially during the early phases of scientific inquiry – the existence of an integrated body of theoretical knowledge cannot be

8.3.2 Programmatic and Prospective Exploration

I introduce the following distinction between "programmatic" and "prospective" exploration with models. On the one hand, programmatic exploration is that which is conducted within the context of a program and does not challenge its core commitments. To put it in different words, programmatic exploration is concerned with exploring commitments at the periphery of a program. In this case, new commitments may be accepted, and new configurations may be attempted, but they affect solely the periphery of the program. On the other hand, prospective exploration is that which challenges the core of the framing program. This challenge comes in the form of acceptance of new commitments, intended to become part of the core of the program or rejection of existing core commitments. As a result, coherence within the core and between peripheral and core commitments may be affected.

Thus, an important question to answer is why scientists would accept commitments that compromise the coherence of their program in its core. To answer this, I need to go back to Chapter 3. In that chapter, I argue that scientists working within research programs typically find themselves in two distinct contexts of justification of commitments. One context is the programmatic context of justification. In this context, I argued that the justification of commitments may be epistemic or practical and follows a coherentism-cum-entrenchment scheme. The other context is the prospective context of justification, whose discussion I postponed for this chapter. The moment has arrived to tackle this issue.

Scientists often accept new commitments that are not part of their ongoing research programs. I refer to these commitments as "commitments-to-be-explored." Commitments-to-be-explored are typically hypotheses, principles, data, phenomena, or methods that are new to a program. They are intended to become the aim of focused study or central tools to be used as part of the research conducted in the program.

assumed, either because such knowledge is not readily available or because it is itself in dispute. The label 'exploratory' is meant to capture just such episodes of scientific inquiry" (15). For this reason, Gelfert (2016) also argues that one of the functions of exploratory models is to serve as "starting points" for further research (84-5). While I agree with these claims, I also consider that exploration can be conducted within well-framed programs. This nuance is one motivation for my distinction between programmatic and prospective exploration.

Commitments-to-be-explored might fit better or worse the configuration of an existing program. Depending on where the commitments-to-be-explored are intended to be placed – i.e., in the periphery or the core of the program – their effects on the overall coherence might be more or less pronounced. Whenever commitments-to-be-explored fit well with the overall configuration of a program, they can be justified via the coherentism-cum-entrenchment scheme (i.e., the programmatic context of justification). This is the case even for those commitments-to-be-explored which do not initially cohere with the program but are intended to play a more peripheral role in the program (recall that coherentism is pondered with entrenchment). Thus, I focus on those commitments-to-be-explored that embody a problem to coherentism-cum-entrenchment, namely those that are intended to be core commitments and do not initially cohere with the core and the program at large in its current configuration.

There is a sense in which the acceptance of initially incoherent (or less than coherent) commitments-to-be-explored is not completely detached from the coherentism-cum-entrenchment scheme. This is because scientists expect commitments-to-be-explored to become coherent with the rest of the program, even though they are initially incoherent. This is what I call "prospective coherence": a form of coherence in a research program that is not present but rather expected to be achieved in the future. This expectation is nurtured by the processual nature of reflective equilibrium and – more generally – the overall piecemeal nature of scientific inquiry. Thus, the fact that a commitment-to-be-explored is initially incoherent with a research program in its core does not mean that the incoming commitment is untenable. It only means that, given the current configuration of the program (i.e., the relations of mutual support among commitments), the incoming commitment does not fit. In any case, prospective coherence does not lend justification to the commitment-to-be-explored at the moment of its acceptance. This is simply because there is no such coherence at the moment of accepting the commitment-to-be-explored (assuming that the commitment-to-be-explored does not initially cohere within the program). A different strategy is required.

My proposed solution is that a form of practical justification is in place when scientists accept a commitment-to-be-explored that does not cohere within their program. Such

practical justification comes in the form of promise, i.e., the potential utility of the commitment-to-be-explored. The potential utility can be cashed out in several forms. There are general forms of potential utility, such as advancing scientific research, developing new lines of research, framing new problems, overcoming degeneration within a program, providing new representational tools, and the like. In particular, prospective coherence can be counted as a general form of potential utility: scientists aim to maximize the coherence in their programs. In addition, there are more specific forms of potential utility, such as delivering potential explanations, qualitative predictions to well-defined problems, add novel descriptions to the program, test the efficiency of a method, and the like. Promise, as a form of practical justification, presumes belief on the utility of an accepted commitment-to-be-explored. Such belief is not blind. It is informed by the past record of the commitment-to-be-explored in contexts different from the host program in question. After all, commitments-to-be-explored might be new to their host programs but that does not mean that they are new altogether. In practice, this amounts to a co-optation of effective commitments across programs.

Co-optation of commitments-to-be-explored is not mere passive adoption. It involves the efforts of scientists to establish connections between the commitment-to-be-explored and their program of origin. These connections often need to be established upon a mostly incoherent contribution of the commitment-to-be-explored to the program that receives them. The gist of this problem is captured in a discussion by Chang (2012). Chang discusses co-option as a form of interaction among programs. He suggests that co-optation involves a kind of incommensurability among programs that has to be overcome (282). I suggest that Chang's incommensurability captures a similar problem to the initial incoherence of commitments-to-be-explored as they arrive to a program. The point can be put this way: commitments-to-be-explored are not ready-made multipurpose entities. They need to be contextualized. Such contextualization involves defining – or at least framing – the role of the commitment-to-be-explained within the host program. Such contextualization does not amount to the affording of immediate and total coherence. It is only a first step towards prospective coherence.

It is not straightforward to answer how scientists co-opt commitments-to-be-explored. There is a myriad of factors to consider. To stress this point, Chang (2012) submits that co-optation in science is "both pervasive and unpredictable" (281). I suspect that delivering an answer to such a problem would amount to providing a putative model for scientific discovery. I am skeptical about the prospects of such enterprise. I conjecture that there is something intricately creative, socially complex, and ultimately unpredictable about actual exploratory work. In any case, this is simply not my enterprise. Having said this, I do not see any reason why proposing a heuristic for exploration – i.e., an enactable suboptimal strategy – must be an ill-fated initiative. In this sense, my distinction between programmatic and prospective exploration is not only intended to be a model of how scientific exploration is conducted. Rather, it is intended as a model of how scientific exploration *could* be conducted promisingly, namely by deciding whether core commitments should be challenged or not in certain research stages.

There is one problem left to tackle, namely the justification process into which commitments-to-be-explored fall in the prospective context. As I have argued, commitments-to-be-explored are justified practically via promise. However, this amounts to a single justification step. If a sceptic keeps asking questions about the acceptance of a commitment-to-be-explored, we soon find ourselves in Agrippa's trilemma territory. If the commitment-to-be-explored coheres with the program, there is no problem with using a coherentist approach. However, if the commitment-to-be-explored does not cohere within the program, a different solution should be provided. My suggestion is that coherentism still functions as a mode of justification of commitments-to-be-explored. However, such coherence is not necessarily to be found at the level of the host program. The idea is that coherence is found at a higher level of organization, in which accepting a commitment-to-be-explored amounts to doing good science. I proceed to develop this idea.

Scientists, in accepting a commitment-to-be-explored, make a judgment about what might retrieve potential utility to their programs, i.e., which are promising commitments to co-opt. These judgments are often not explicated by those scientists delivering the judgment,

i.e., they are tacit.[99] They are often described as *having a feeling for*, and they are referred to with terms such as "intuitions" or "hunches". A more technical approach in the literature to deal with this issue is that of "tacit knowledge".

There are two contrasting approaches to tacit knowledge that are presented in the work of two major contributors to the debate. On the one hand, Polanyi (2009) argues that tacit knowledge that supports expert judgments is grounded on the level of the individual. For Polanyi, there is an element of the personality of a scientist that is inextricable from scientific discovery (24-5). On the other hand, Collins (2010) suggests that much of what Polanyi describes as personal knowledge can be cashed out as a form of collective tacit knowledge. In this view, what Polanyi often refers to as "intuitions" can be gained through practice and socialization (149). I consider that Collins's program accounts for various of Polanyi's concerns in a less mysterious fashion. Accordingly, I follow his proposal.

Scientists who propose to adopt a new – and even incoherent – commitment-to-be-explored in their programs exert a kind of tacit knowledge in their judgments. Such tacit knowledge is collective in the sense that choosing a commitment-to-be-explored must align with what is *collectively* considered as good scientific practice. It is at this level that coherence can be found. And, accordingly, the co-optation of a commitment-to-be-explored can be justified via coherentism at this level. To be sure, it is not straightforward to say what the standards for good science are. Arguably, these standards are continuously being modified by practicing scientists and, depending on the context, different standards might apply. Nonetheless, I have a few reflections on this issue.

Using Collins and Evans (2007)'s terminology, the tacit knowledge employed in co-opting a commitment-to-be-explored can be characterized as *ubiquitous* and *specialist* tacit knowledge. Roughly, in order to co-opt a commitment from other program, scientists must possess some degree of acquaintance with what is going on in other programs. Such

[99] These judgments might be not explicated as a matter of contingency. However, some judgments are not explicated because they are not explicable in some definite sense. For example, Collins (2010) provides a list of "cannots" that he uses to qualify what he means by saying that something (in particular, a judgment) *cannot* be made explicit (89, Table 5).

acquaintance can go from basic to highly refined. For example, a scientist might have "beer-mat" knowledge of another program, a kind of ubiquitous tacit knowledge that reaches high levels of succinctness and simplicity. Or, for example, a scientist might have interactional expertise, a kind of specialist tacit knowledge that amounts to fluency in the language of a program in which one is not a practitioner. It is difficult to say where the standards of good scientific co-optation stand in this spectrum. Arguably, good scientists – at some point – rely on their ubiquitous knowledge. In this sense, even beer-mat knowledge cannot be discarded, especially in early exploratory stages with thought experiments and toy models. Certainly, scientists possessing interactional expertise or at least primary source knowledge (i.e., knowledge derived from reading scientific papers), is a desideratum in good science.

In any case, none of this acquaintance with other programs serves any purpose to a scientist's home program if she does not exert what Collins and Evans refer to as "contributory expertise." This is a kind of specialist tacit knowledge that enables scientists to do things in their domain of expertise. In practice, the domain of expertise amounts to a scientist' home program. Such expertise is gained by means of engaging in the practices of a program. I suggest that this is the kind of tacit knowledge that is required to materialize the promise of a co-opted commitment.

I point at two senses in which contributory expertise is required. First, the co-opted commitment must be contextualized. As I explained above, this amounts to establishing the role that a commitment-to-be-explored will have in its host program. Such role can hardly be established if one does not have expertise in terms of the practices within the program. Second, contributory expertise converts the potential utility of a commitment-to-be explored – its promise – into actual utility. That is, by knowing the practices of a program, scientists can design and conduct meaningful and effective exploratory work – typically experimentation and modeling – involving the commitment-to-be-explored.

To close the theoretical discussion on programmatic and prospective explorations, I would like to express some qualifications. As I argued in Chapter 3, there is a significant amount of idealizations involved in the individuation of research programs. There are also various idealizations involved in representing these programs as a binary core-periphery

structure. The reality of science and scientific practices is messier. For these reasons, the programmatic-prospective scheme should be regarded merely as a model of scientific exploration, not as a complete and accurate description of it. I estimate that such a model does adequate descriptive work in the BK and OFC cases (as I attempt to show below). However, its representational adequacy must be tested in other contexts. In addition, I suggest that the model could be used as a heuristic for exploration. My assessment of the BK and OFC cases provides a proof of possibility for this claim.

I proceed to test my ideas on exploration in the context of my case studies. I consider first the BK case. BK argue that their project is significantly different from what was the core aim of model and theoretical seismology back then, namely the study of the propagation of seismic waves across the Earth. BK's problem is a different one: it is the study of the occurrence of earthquakes. This is what they refer to as seismicity (S). S is a commitment-to-be-explored in the form of an aim. And, importantly, this is not some sort of subsidiary aim: it is intended to be placed at the core of BK's enterprise. Prima facie, there is nothing incoherent about placing S at the core of seismology. In fact, most seismologists working today would regard S as a problem in seismology. This is arguably what BK were trying to do: to extend the scope of the problems of seismology. BK were in a privileged position to attempt this. They were well-reputed seismologists, who had done programmatic research on the propagation of seismic waves across continuous media. In addition, they had been conducting pioneering research on the occurrence of earthquakes for some time. In fact, their 1964 paper "Body Force Equivalents for Seismic Dislocations" has been described as playing "a critical role in the development of the theory of the earthquake source" (Pujol, 2003). In particular, BK's "representation theorem" has been referred to as "the first principle in modern seismology" (Doel, 1990).

The point I want to emphasize is that, as a matter of contingency, S was not an aim – or, at least, not a core one – in seismology at the time BK conducted their research. As a result, the overall constellation of commitments in seismology was not built and configured in a way that allowed S to be adopted straightforwardly. In other words, the problem arises when trying to study S with the available methods of seismology, namely theoretical and model seismology. As BK argue, theoretical seismology is not fit to address the problem of seismicity:

there is no physical theory of earthquake occurrence in analytic form. Given that S cannot be dealt with by means of theoretical seismology, model seismology is out of the question. (Recall that model seismology amounts to constructing an analogue computer that solves the problem of theoretical seismology.)

And, nonetheless, BK intend to use theoretical and model approaches to tackle S – hence the title of their paper: Model and Theoretical Seismicity. BK's solution is to construe S in such a way that makes it susceptible to being treated by means of theoretical and model approaches. This is where the acceptance of P1 and P2 plays a central role. Recall that P1 is the suitability of a discrete system to represent and simulate the behavior of a continuum medium. And P2 is the sufficiency of friction as a causal factor to account for the main features of the statistics of naturally occurring earthquakes. Up to their paper, established qualitative theories of earthquake occurrence "have perforce been made elaborate because of the large number of features they have been called upon to explain" (342). BK accept P1 and P2 as commitments-to-be-explored to construe S in a way susceptible to being tackled via theoretical and model approaches. More explicitly, BK focus on friction as an isolated causal factor in geological faults, and they represent the fault as a discrete system. This construal of S allows BK to implement theoretical and model approaches to S: they provide a system of ordinary differential equations, susceptible to being solved via a computer program. Given their central role in BK's enterprise, and subsequent ubiquitous impact in their research, P1 and P2 are core commitments, together with S.

Acceptance of S, P1 and P2 as commitments-to-be-explored is justified practically. S is BK's aim, their main motivation. And P1 and P2 are means to study S. In addition, the justification process for accepting S, P1 and P2 appeals to the coherence of these commitments-to-be-explored with good scientific practices. This amounts to saying that S is a well-defined and significant problem to be studied scientifically. And the co-optation of P1 and P2 is informed by specialist (and arguably also ubiquitous) tacit knowledge. P1 is a technique that had been applied in other contexts to attain the description of a continuous dynamic system via ordinary differential equations. And P2 is based on a series of qualitative models on the generation of earthquakes in geological faults.

The core of BK's program – i.e., S, P1, and P2 – is a coherent set of commitments. But, from the point of view of traditional seismological studies (back then) this set of commitments did not fit well into the configuration of aims, assumptions, and practices of seismology. BK were advancing a new problem, which was tackled via unconventional means. In this sense, BK's exploration is prospective: it embodies an assemblage of principles for the construction and interpretation of their model that challenges core tenets in model and theoretical seismology at the time. It is clear that model and theoretical seismology are the programs to which BK are subscribing: BK frame their research as a response, or extension, to model and theoretical seismology. In addition, the paper is submitted to the Bulletin of the Seismological Society of America. Just to be clear, S, P1 and P2 could have been part of these programs in seismology. There is nothing untenable about that. It just happens to be the case that they were not part of the coherent set of commitments built into the core of seismology at that point.

A notorious outcome of BK's exploratory motivations is the design of the spring-block model, a simple mechanism which departs from previous efforts in the field. It is by means of this model that BK put themselves in the position to think about earthquakes in a mechanistic way. By constructing and manipulating a discrete and modular mechanism in the laboratory, or describing it with Newtonian physics and simulating it, BK expect to learn facts that may be imputed to seismic faults, even if only with modal qualifications. Thus, the BK case is an example of the exploratory benefits of understanding differently (in this case, differently from established seismological approaches at the time).

In the OFC case, a similar analysis can be offered. OFC's exploratory efforts can be characterized as specific and indirect, and involving isolations, roughly for the same reasons presented for the BK case. Indeed, OFC explore the SOC behavior of a specific target, namely earthquakes in seismic faults (i.e., specific exploration). This exploration is conducted not in actual seismic faults but by means of a model (i.e., indirect exploration). And the OFC model, like the BK model, focuses on isolated causal factors, namely the dynamics of friction (isolation).

The evaluation of OFC's exploratory strategies as programmatic or prospective has some relevant nuances. Hence, I propose to dissect the analysis. The reason is that OFC frame their research as a response to two distinct programs, namely BK's program of earthquake occurrence and the SOC program. The former had established a mechanistic understanding of earthquakes. The latter, for the most part, had dealt with mathematical modeling of various systems in the form of cellular automata simulations. Thus, in order to evaluate whether OFC's exploration is programmatic or prospective, one has to establish which program is the one from which the evaluation is conducted.

From the perspective of the BK program, I suggest that the OFC model is an exploratory model in a prospective sense. The decisive reason is that OFC engage with the spring-block metaphor proposed by BK, but they re-interpret it as a mathematical entity.[100] This interpretation is guided by mathematical explanatory commitments, which are core commitments in the SOC program. However, these mathematical commitments challenge BK's core mechanistic commitments.[101] Still, from the perspective of the SOC program, the OFC model performs programmatic exploration. This is manifest in OFC's decision to construct a model that does not challenge the modeling tradition in the SOC program and adopts its core commitments, particularly its mathematical explanatory ones. In simple words, what OFC do is introduce the spring-block model of earthquakes to the SOC program. Such introduction stays in line with the SOC program but requires a re-interpretation of the spring-block system in terms of "challenging" mathematical explanatory commitments.[102]

---

[100] Metaphors have been previously discussed as material for scientific exploration (e.g., Bailer-Jones, 2010: 118-9). However, it is not clear how this exploratory function takes place. For example, Johnson (2002) argues that "each metaphor will carry its own distinctive matrix of values that guide the research that flows from the metaphor" (14). Thus, using different metaphors affords different value stances from which to reason about a target. The spring-block picture, as a metaphor, shows that metaphors not only embody their own distinctive matrix of commitments, but can also be re-interpreted in light of a different matrix of commitments. In my case studies, the spring-block metaphor affords mechanistic understanding to BK and mathematical understanding to OFC. Such re-interpretation is an exploratory move. It provides new content and metacontent to a metaphor in light of which scientists can reason afresh about a target.

[101] It is worth noticing that OFC take their model to be "equivalent" to a two-dimensional version of BK's spring-block model (Christensen and Olami, 1992a: 8729). I consider this characterization unfortunate because it is their very differences in terms of content that allow OFC to think about earthquakes in a more abstract way. Certainly, OFC and BK resort to the same metaphor, but the metaphor is non-representationally interpreted in contrasting fashions.

[102] Just to be clear, OFC do challenge previous assumptions within the SOC program, namely the assumption of requiring conservative regimes for SOC behavior and discrete state variables. However, these assumptions are not core commitments of the SOC program, but peripheral to it. In challenging them, OFC overall assumptions

Thus, I conclude that understanding differently plays an exploratory role. The BK and OFC cases show two distinct ways this can happen. BK construct a novel model based on newly assembled principles, that is, the spring-block system. OFC reinterpret the spring-block model under different explanatory commitments, thus achieving a different understanding. In both cases, understanding differently – i.e., using models with distinct non-representational content – led to discovery of facts, even if modal ones.

8.4 Précis

The main theoretical theses of this chapter are:

- Thesis 1: Understandingly differently plays an exploratory role. By understanding differently, different kind of model explanations can be provided, even modally qualified.
- Thesis 2: Exploration can be conducted programmatically or prospectively. Programmatic exploration does not challenge the core of a program, while prospective exploration does.

The main lessons learnt from my case studies are:

- EU and OU are attributed to BK and OFC.
- The non-representational content of BK and OFC's state of understanding is different. BK possess mechanistic understanding. OFC possess mathematical understanding. In other words, BK and OFC understand the occurrence of earthquakes differently.
- BK engage in prospective exploration with their model.
- From the perspective of BK's program, OFC engage in prospective exploration by re-interpreting the spring-block model with challenging core commitments of a mathematical sort. From the perspective of the SOC program, OFC engage in programmatic explanation: they introduce a new cellular-automaton model.

***

remain coherent with SOC's main explanatory commitments, modeling practices, and core theoretical tenets. For more details, see the discussion in Section 3.3.3.

9 Epilogue: In Defense of Explanatory Pluralism

Understanding differently, by means of accepting distinct explanatory commitments, plays an important role in scientific research, especially in exploratory stages. The adoption of these different commitments allows scientists to engage in modeling practices that shed new light upon ongoing problems. As a consequence, this dissertation can be read as providing evidence for explanatory pluralism. This does not only mean that this dissertation shows that there is a plurality of explanatory practices. It also means that this plurality has merits, as I have shown in previous chapters.

There is an old tradition of explanatory pluralism in philosophy of science, whose origins can be traced back to Aristotle and his four causes. More recently, philosophers of science still advocate this position. For example, Lipton (2008) argues that "[w]hen it comes to scientific explanation, we should be pluralists" (2008: 124). Pincock (2018) argues that "the diversity of explanations found in scientific practice found in scientific practice mandates some form of explanatory pluralism." And Mantzavinos (2016) advocates a form of explanatory pluralism, not merely as an accurate description of scientific practice, but also as a valuable stance that enables the solution of problems. Using Mantzavinos's terminology (2016), it is valuable to play different explanatory games.

To finish this dissertation, I would like to briefly comment on an enlightening feature of Mantzavinos's term 'explanatory games'. Explanatory games are games not only in the sense of having their own rules. They are also games in the sense of soliciting a "playful" attitude from scientists (thanks to Martin Kusch for suggesting this to me). Playfulness is key to advancing scientific research and, in particular, scientific explanations. As Mantzavinos puts it: "Explanations emerge in the process of playing the game" (35). And playfulness enables scientists to play different games. This way, new explanations can be crafted or, as Mantzavinos would put it, new solutions to problems can be found.

Certainly, playfulness may have its limits, which are often embodied as intersubjective standards. In fact, I have shown in several passages of this dissertation how programmatic

research poses some restraints upon playfulness (e.g., justificatory schemes). Other forms of moderation may be required in responding to intersubjective standards of responsible research. It is relevant to discuss how to combine these two attitudes – the playful and the responsible – in a way that fosters better science. This, however, is a matter that I leave for my next research project.

***

10 References

Achinstein, P. (1983). The Nature of Explanation. New York: Oxford University Press.

Alexandrova, A. & Northcott, R. (2013). It's just a feeling: why economic models do not explain. Journal of Economic Methodology, 20(3): 262-267.

Ankeny, R. & Leonelli, S. (2016). Repertoires: A post-Kuhnian perspective on scientific change and collaborative research. Studies in History and Philosophy of Science, 60: 18-28.

Avnir, D., Biham, O., Lidar, D. & Malcai, O. (1998). Is the Geometry of Nature Fractal? Science, 279(5347): 39-40.

Bailer-Jones, D. M. (2009). Scientific models in philosophy of science. Pittsburgh: University of Pittsburgh Press.

Bak, P. (1996). How Nature Works: The Science of Self-Organized Criticality. New York: Copernicus.

Bak, P. & Chen, K. (1989). The physics of fractals. Physica D, 38: 5-12.

Bak, P. & Chen, K. (1991). Self-Organized Criticality. Scientific American, 264(1): 46-53.

Bak, P. & Tang, C. (1989). Earthquakes as a Self-Organized Critical Phenomenon. Journal of Geophysical Research, 94(B11): 15,635-7.

Bak, P., Tang, C. & Wiesenfeld, K. (1987). Self-Organized Criticality: An Explanation of 1/f Noise. Physical Review Letters, 59(4): 381-4.

Bak, P., Tang, C. & Wiesenfeld, K. (1988). Self-Organized Criticality. Physical Review A 38: 364-374.

Barberousse, A., Franceschelli, S. & Imbert, C. (2007). Cellular automata, modeling, and computation. PhilSci Archive.

Bartels, A. (2006). Defending the structural concept of representation. Theoria, 21(1): 7-19.

Båth, M. (1958). The energies of seismic body waves and surface waves. Contributions in Geophysics in honor of Beno Gutenberg. London: Pergamon Press. pp: 1-16.

Batterman, R. & Rice, C. (2014). Minimal model explanations. Philosophy of Science, 81: 349-76.

Batterman, R. W. (2002). The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence. Oxford University Press.

Baumberger, C., Beisbart, C. & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. Grimm, C. Baumberger & S. Ammon (Eds.) “Explaining Understanding” (pp. 1-34). New York: Routledge.

Bechtel, W., & Richardson, R. C. (2010). Discovering complexity: Decomposition and localization as strategies in scientific research. MIT press.

Bedau, M. (1997). Weak emergence. Noûs, 31, Supplement: Philosophical Perspectives, 11, Mind, Causation, and World: 375-399

Berlyne, D. (1960). Conflict, arousal, and curiosity. New York: McGraw-Hill

Berto, F. & Tagliabue, J. (2017). Cellular Automata. The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/cellular-automata/>.

Blatter, J. (2008). Case study. In Given, L. M. (Ed.). The Sage encyclopedia of qualitative research methods. Sage publications.

Bokulich, A. (2011). How scientific models can explain. Synthese, 180: 33-45.

Bokulich, A. (2012). Distinguishing explanatory from nonexplanatory fictions. Philosophy of Science, 79(5), 725-737.

Bokulich, A. (2014). How the Tiger Bush got its Stripes: 'How Possibly' vs. 'How Actually' model explanations. The Monist, 97(3): 321-38.

Bokulich, A. (2018). Representing and explaining: The eikonic conception of scientific explanation. Philosophy of Science, 85(5), 793-805.

Brace, W. & Byerlee, J. (1966). Stick-slip as a mechanism for earthquakes. Science, New Series, 153(3739), 990-992.

Brandon, R. (1990). Adaptation and Environment. Princeton: Princeton University Press.

Bridgman, P. (1950). Reflections of a Physicist. New York: Philosophical Library.

Brown, S. R., Scholz, C. H., & Rundle, J. B. (1991). A simplified spring-block model of earthquakes. Geophysical Research Letters, 18(2), 215-218.

Burridge, R., & Knopoff, L. (1964). Body force equivalents for seismic dislocations. Bulletin of the Seismological Society of America, 54(6A), 1875-1888.

Burridge, R., & Knopoff, L. (1967). Model and theoretical seismicity. Bulletin of the seismological society of America, 57(3), 341-371.

Carlson, J. M., & Langer, J. S. (1989). Properties of earthquakes generated by fault dynamics. Physical Review Letters, 62(22), 2632.

Carnap, R. (1950). Empiricism, semantics and ontology. Revue Internationale de Philosophie 4, 20-40.

Cartwright, N. (1983). How the laws of physics lie. Oxford: Oxford University Press.

Caruso, F., Latora, V., Pluchino, A., Rapisarda, A. & Tadic, B. (2006). Olami-Feder-Christensen model on different networks. European Physical Journal, 50: 243-7.

Chang, H. (2012). Is Water H2O? Evidence, Realism and Pluralism. London: Springer.

Chang, H. (2017). VI—Operational Coherence as the Source of Truth. Proceedings of the Aristotelian Society, 117(2): 103-22.

Christensen, K. & Olami, Z. (1992a). Scaling, phase transitions, and nonuniversality in self-organized critical cellular-automaton model. Physical Review A, 46(4): 1829-1838.

Christensen, K. & Olami, Z. (1992b). Variation of the Gutenberg-Richter b values and nontrivial temporal correlations in a spring-block model for earthquakes. Journal of Geophysical Research, 97(B6), 8729-8735.

Clancy, I. & Corcoran, D. (2006). Burridge-Knopoff model: Exploration of dynamic phases. Physical Review E 73, 046115.

Colletini, C., Niemeijer, A., Viti, C. & Marone, C. (2009). Fault zone fabric and fault weakness. Nature, 462: 907-11.

Collins, H. (2010). Tacit and explicit knowledge. Chicago: The University of Chicago Press.

Collins, H. & Evans, R. (2007). Rethinking Expertise. Chicago: University of Chicago Press.

Contessa, G. (2007). Scientific Representation, Interpretation, and Surrogative Reasoning. Philosophy of Science, 74: 48–68.

Costagliola, G., Bosia, F. & Pugno, N. (2017) Hierarchical Spring-Block Model for Multiscale Friction Problems. ACS Biomaterials Science & Engineering, 3(11): 2845-52

Cowling, S. (2017). Resemblance. Philosophy Compass, 12(4): e12401

Craver, C. (2006). When mechanistic models explain. Synthese, 153: 355–376.

Craver, C. (2014). The Ontic Account of Scientific Explanation. In Kaiser, M.I., Scholz, O.R., Plenge, D. & Hüttemann, A. (Eds.) "Explanation in the Special Sciences: The Case of Biology and History" (Ch. 2, pp. 27-52), Synthese Library, 367. Dordrecht: Springer.

Craver, C. F. (2007). Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Oxford University Press.

Cuffaro, M. (2015) - How-Possibly Explanations in (Quantum) Computer Science. Philosophy of Science, 82: 737–748.

Curie, A. (2015). Philosophy of Science and the Curse of the Case Study. In Daly, C. (Ed.). The Palgrave handbook of philosophical methods. Springer.

Dascal, M. (2003). Interpretation and Understanding. Amsterdam: John Benjamins Publishing Company.

de Regt, H. & Dieks, D. (2005). A Contextual Approach to Scientific Understanding. Synthese, 144: 137-170.

de Regt, H. & Gijsbers, V. (2017). How false theories can yield genuine understanding. In S. Grimm, C. Baumberger & S. Ammon (Eds.) "Explaining Understanding" (pp. 50-75). New York: Routledge.

de Regt, H. W. (2017). Understanding scientific understanding. Oxford University Press.

de Regt, H. W., Leonelli, S., & Eigner, K. (Eds.). (2009). Scientific understanding: Philosophical perspectives. University of Pittsburgh Press.

de Sousa Vieira. (1995). Chaos in a simple spring-block system. Physics Letters A, 198(5-6): 407-14.

Dhar, D. (1990). Self-organized critical state of sandpile automaton models. Physical Review Letters 64(14): 1613-6.

Doel, R. (1990). Interview with Dr. Leon Knopoff (Transcript). American Institute of Physics.

Douglas, H. (2016). Values in science. In P. Humphreys (Ed.) "The Oxford Handbook of Philosophy of Science" (24pp) New York: Oxford University Press.

Dowe, P. (1992). Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory. Philosophy of Science, 59, 195-216.

Dray, W. (1957). Law and Explanation in History. Oxford: Oxford University Press.

Droysen, J. (1868). Grundriss der Historik. Leipzig: Veit.

Duran, J. & Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. Minds and Machines, 28: 645-666.

Elgin, C. (1996). Considered Judgement. Princeton, N.J.: Princeton University Press.

Elgin, C. (2009). Exemplification, idealization and understanding. In M. Suárez (Ed.) "Fictions in Science" (pp. 77-90). New York: Taylor & Francis Group.

Elgin, C. (2017). True enough. MIT Press.

Elgin, M. & Sober, E. (2002). Cartwright on explanation and idealization. Erkenntnis, 57: 441–450.

Faye, J. (2014). The Nature of Scientific Thinking: On Interpretation, Explanation and Understanding. New York: Palgrave Macmillan.

Feder, H. & Feder, J. (1991). Self-organized criticality in a stick-slip process. Physical Review Letters, 66(20): 2669–72.

Ferguson, C., Klein, W. & Rundle, J. (1998). Long Range Earthquake Fault Models. Computers in Physics, 12(1): 34-40.

Fischer, B. & Leon F. 2017. Modal Epistemology After Rationalism. Cham: Springer

Forber, P. 2010. Confirmation and explaining how possible. Studies in History and Philosophy of Biological and Biomedical Sciences, 41: 32-40.

French, S. (2003). A model-theoretic account of representation. Philosophy of Science, 70(5): 1472-1583.

Friedman, M. (1974). Explanation and Scientific Understanding. Journal of Philosophy, 71: 5–19.

Friedman, M. (2001). Dynamics of Reason: The 1999 Kant Lecture at Stanford University. Stanford: CSLI Publications.

Frigg, R. (2003a). Self-organised criticality—what it is and what it isn't. Studies in History and Philosophy of Science Part A, 34(3), 613-632.

Frigg, R. (2003b), Re-presenting Scientific Representation. PhD. Dissertation. London: London School of Economics.

Frigg, R. & Hartmann, S. (2020). "Models in Science", The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/models-science/>.

Frigg, R. & Nguyen, J. (2016). The Fiction View of Models Reloaded. The Monist, 99: 225–242.

Frigg, R., & Nguyen, J. (2018). The turn of the valve: representing with material models. European Journal for Philosophy of Science, 8(2), 205-224.

Gelfert, A. (2016). How to do science with models: A philosophical primer. Cham: Springer.

Gelfert, A. (2017). The Ontology of Models. In L. Magnani & T. Bertolotti, T. (eds.) "Springer Handbook of Model-based Science" (pp. 5-24). Cham: Springer.

Gelfert, A. (2019). Probing possibilities: Toy models, minimal models, and exploratory models. In: Nepomuceno-Fernández Á., Magnani L., Salguero-Lamillar F., Barés-Gómez C., Fontaine M. (eds) Model-Based Reasoning in Science and Technology. MBR 2018. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 49. Springer, Cham.

Gerring, J. (2007). Case study research: Principles and practices. Cambridge university press.

Giere, R. (1988). Explaining Science. Chicago: The University of Chicago Press

Giere, R. (2006). Scientific Perspectivism. Chicago: The University of Chicago Press.

Glanzberg, M. (2018). "Truth", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/truth/>.

Glennan, S. (1996). Mechanisms and the Nature of Causation. Erkenntnis, 44, 49-71.

Glennan, S. (2017). The new mechanical philosophy. Oxford University Press.

Glennan, S., & Illari, P. (Eds.). (2017). The Routledge handbook of mechanisms and mechanical philosophy. Taylor & Francis.

Goldberg, N. (2009). Historicism, Entrenchment, and Conventionalism. Journal for General Philosophy of Science, 40(2): 259-276.

Gombrich, E. (1960). Art and Illusion: A study in the psychology of pictorial representation. Princeton: Princeton University Press.

Goodman, N. (1955). Fact, Fiction, and Forecast. Cambridge: Harvard University Press.

Goodman, N. (1976). Languages of Art: An Approach to a Theory of Symbols. Indianapolis: Bobbs-Merrill.

Grassberger, P. (1994). Efficient large-scale simulations of a uniformly driven system. Physical Review E, 49(3): 2436-44.

Grimm, S. (2017). Understanding and transparency. In S. Grimm, C. Baumberger & S. Ammon (Eds.) "Explaining Understanding" (pp. 212-119). New York: Routledge.

Grimm, S. R., Baumberger, C., & Ammon, S. (Eds.). (2017). Explaining understanding: New perspectives from epistemology and philosophy of science. Taylor & Francis.

Grinstein, G., D.-H. Lee, & Sachdev, S. (1990). Conservation laws, anisotropy, and 'self-organized criticality' in noisy nonequilibrium systems. Physical Review Letters, 66(2): 1927-30.

Gros, C. (2015). Complex and Adaptative Dynamical Systems: A primer. Cham: Springer.

Guala, F. (2003). Experimental Localism and External Validity. Philosophy of Science, 70(5): 1195-205.

Hempel, C. (1965). Aspects of scientific explanation. In C. Hempel C (ed.) "Aspects of scientific explanation and other essays in the philosophy of science" (pp. 331-496). New York: Free Press.

Hempel, C. & Oppenheimer, P. (1948). Studies in the logic of explanation. Philosophy of Science, 15: 135-75.

Hergarten, S. (2002). Self organized criticality in earth systems (Vol. 2, No. 2). Berlin: Springer.

Hughes, R. (1997). Models and Representation. Philosophy of Science, 64: S325-S336

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. Synthese, 169(3), 615-626.

Hwa, T. & Kardar, M. (1989). Dissipative transport in open systems: an investigation of self-organized criticality. Physical Review Letters 62(16): 1813-6.

Illari, P. (2013). Mechanistic Explanation: Integrating the Ontic and Epistemic. Erkenntnis, 78(Supplement 2), 237-255.

Illari, P. (2019). Mechanisms, models and laws in understanding supernovae. Journal for General Philosophy of Science, 50: 63-84.

Illari, P. M. & Williamson, J. (2012). What is a Mechanism? Thinking about Mechanisms across the Sciences. European Journal for Philosophy of Science, 2, 119-35.

Járai-Szabo, F., Sándor, B. & Néda, Z. (2010). Spring-block model for a single-lane highway traffic. Central European Journal of Physics, 9(4): 1002-9.

Jebeile, J. & Graham Kennedy, A. (2015). Explaining with models: The role of idealizations. International Studies in the Philosophy of Science, 29(4): 383-92.

Jensen, H. J. (1998). Self-organized criticality: emergent complex behavior in physical and biological systems (Vol. 10). Cambridge university press.

Johansen, A., Dimon, P, Ellegaard, C., Larsen, J., & Rugh, H. (1993). Dynamic phases in a spring-block system. Physical Review E, 48(6): 4779-90.

Johnson, M. (2002). Metaphor-based values in scientific models. In L. Magnani & N. Nersessian (eds.) "Model-based reasoning: Science, technology, values" (pp. 1-21). New York: Kluwer Academic/Plenum Publishers.

Julien, H. (2008). Content analysis. In Given, L. M. (Ed.). (2008). The Sage encyclopedia of qualitative research methods. Sage publications.

Kagan, Y. Y. (2014). Earthquakes: models, statistics, testable forecasts. John Wiley & Sons.

Kaplan, D., & Craver, C. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. Philosophy of Science, 78: 601–627.

Khalifa, K. (2013). Is understanding explanatory or objectual? Synthese, 190(6): 1153-1171.

Khalifa, K. (2017). Understanding, explanation, and scientific knowledge. Cambridge University Press.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon W (eds.) "Scientific explanation" (pp. 410-505). Minneapolis: University of Minnesota Press.

Kment, B. (2017). "Varieties of Modality", The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/modality-varieties/>.

Knuuttila, T. (2005). Models as epistemic artefacts: Toward a non-representationalist account of scientific representation. Philosophical Studies from the University of Helsinki.

Knuuttila, T. (2010). Some Consequences of the Pragmatist Approach to Representation: Decoupling the Model-Target Dyad and Indirect Reasoning. In M. Suárez, M. D., & M. Rédei (Eds.), EPSA epistomology and methodology of science: Launch of the European philosophy of Science Association (pp. 139-148). (EPSA; Vol. 1). Dordrecht: Springer.

Knuuttila, T. & Morgan, M. (2019). Deidealization: No easy revearsals. Philosophy of Science, 86(4): 641-661.

Knuuttila, T. & Voutilainen, A. (2003). A Parser as an Epistemic Artifact: A Material View on Models. Philosophy of Science, 70(5): 1484-1495.

Kuhn, T. (1996). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Kvanvig, J. (2009). The value of understanding. In A. Haddock, A. Millar, & D. Pritchard (Eds.) "Epistemic value" (pp.95-111). Oxford: Oxford University Press.

Kvanvig, J. L. (2003). The value of knowledge and the pursuit of understanding. Cambridge University Press.

Lakatos, I. (1978). The Methodology of Scientific Research Programmes. In J. Worrall & G. Currie (eds.) Philosophical Papers, Vol. 1. Cambridge: Cambridge University Press.

Lawler, I. (2019). Scientific understanding and felicitous legitimate falsehoods. Synthese (online first).

Le Bihan, S. (2017). Enlightening falsehoods: A modal view of scientific understanding. In S. Grimm, C. Baumberger & S. Ammon (Eds.) "Explaining Understanding" (pp. 111-135). New York: Routledge.

Levy, A. (2013). Three kinds of new mechanism. Biology & Philosophy, 28: 99-114.

Lipton, P. (2008), 'CP Laws, Reduction and Explanatory Pluralism'. In J. Hohwy and J. Kallerstrup (eds.), "Being Reduced: New Essays on Reduction, Explanation and Causation" (pp. 115-25) Oxford: Oxford University Press.

Lipton, P. (2009). Understanding without Explanation. In H. de Regt, S. Leonelli & K. Eigner (eds.) "Scientific Understanding: Philosophical Perspectives" (pp. 43-63). Pittsburgh: University of Pittsburgh Press.

Machamer, P. K., Darden, L. & Craver, C. F. (2000). Thinking about Mechanisms. Philosophy of Science, 67, 1-25.

Magnani, L., & Bertolotti, T. (Eds.). (2017). Springer handbook of model-based science. Springer.

Magnani, L., & Nersessian, N. J. (Eds.). (2002). Model-based reasoning: Science, technology, values. Springer Science & Business Media.

Magnani, L., Nersessian, N., & Thagard, P. (Eds.). (1999). Model-based reasoning in scientific discovery. Springer Science & Business Media.

Mäki, U. (2009a). Realistic realism about unrealistic models. In H. Kincaid & D. Ross (Eds.) "The Oxford Handbook of Philosophy of Economics" (pp. 68-98). Oxford: Oxford University Press.

Mäki, U. (2009b). MISSing the world. Models as Isolations and credible surrogate systems. Erkenntnis, 70: 29-43.

Mandelbrot, B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science, New Series, 156(3775): 636-8.

Manna, S., Kiss, L. & Kertész, J. (1990). Cascades and self-organized criticality. Journal of Statistical Physics, 61(3/4): 923–932.

Mantzavinos, C. (2016). Explanatory Pluralism. Cambridge: Cambridge University Press.

Margolis, Eric and Laurence, Stephen, "Concepts", The Stanford Encyclopedia of Philosophy (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/concepts/>.

Massimi, M. (2018). Perspectival modeling. Philosophy of Science, 85(3), 335-359.

McMullin, E. (1982). “Values in Science”, PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1982, 3–28.

McMullin, E. (1985). Galilean idealization. Studies in History and Philosophy of Science, 16, 247–273.

Morgan, M. S., & Morrison, M. (1999). Models as mediators (p. 347). Cambridge: Cambridge University Press.

Morrison, M. (2015). Reconstructing reality: Models, mathematics, and simulations. Oxford Studies in Philosophy of Science.

Nakanishi, H. (1990). Cellular-automaton model of earthquakes with deterministic dynamics. Physical Review A, 41(12), 7086.

Olami, Z., Feder, H. J. S., & Christensen, K. (1992). Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Physical review letters, 68(8): 1244-1248

Otsuka, M. (1972). A Simulation of Earthquake Occurrence. Physics of the Earth and Planetary Interiors, 6: 311-5.

Parker, W. (2009). Does matter really matter? Computer simulations, experiments, and materiality. Synthese, 169: 483-96.

Parker, W. (2014). Simulation and understanding in the study of weather and climate. Perspectives on Science, 22(3): 336-56.

Persson, J. (2012). Three conceptions of explaining how possibly—and one reductive account. In Henk W. de Regt, Stephan Hartmann, & Samir Okasha (eds.) “EPSA Philosophy of Science: Amsterdam 2009” (pp. 275–286). Dordrecht: Springer

Pincock, C. (2012). Mathematics and scientific representation. Oxford University Press.

Pincock, C. (2018). Accommodating Explanatory Pluralism. In A. Reutlinger & J. Saatsi (Eds.). “Explanation beyond causation: philosophical perspectives on non-causal explanations” (pp. 39-56). Oxford: Oxford University Press.

Polanyi, M. (2009). The tacit dimension. London: University of Chicago Press

Portides, D. (2017). Models and theories. In L. Magnani & T. Bertolotti (Eds.) “Springer Handbook of Model-based Science” (pp: 25-48). Cham: Springer.

Potochnik, A. (2017). Idealization and the Aims of Science. University of Chicago Press.

Pruessner, G. (2012). Self-organised criticality: theory, models and characterisation. Cambridge University Press.

Pujol, J. (2003). Introduction to the Classic Paper: Body Force Equivalents for Seismic Dislocations, by R. Burridge and L. Knopoff. Seismological Research Letters, 74(2): 153

Putelat, T., Dawes, J. & Willis, J. (2010). Regimes of frictional sliding of a spring–block system. Journal of the Mechanics and Physics of Solids, 58(1): 27-53

Quine, W. V. (1986). Philosophy of logic. Harvard University Press.

Race, R. (2008). Literature Review. In Given, L. M. (Ed.). (2008). The Sage encyclopedia of qualitative research methods. Sage publications.

Rawls, J. (1999), A Theory of Justice. Cambridge, MA: Harvard University Press.

Reiss, J. & Sprenger, J. (2014). Scientific Objectivity. The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/entries/scientific-objectivity/>.

Resnik, D. (1991). How-possibly explanations in biology. Acta Biotheoretica, 39: 141-9.

Reutlinger, A., & Saatsi, J. (Eds.). (2018). Explanation beyond causation: philosophical perspectives on non-causal explanations. Oxford University Press.

Reutlinger, A., Hangleiter, D. & Hartmann, S. (2018). Understanding (with) Toy Models. British Journal for the Philosophy of Science, 69(4): 1069–1099.

Reydon, T. (2012). Discussion: How-Possibly Explanations as Genuine Explanations and Helpful Heuristics: A Comment on Forber. Studies in History and Philosophy of Biological and Biomedical Sciences, 43(1): 302-10.

Rice, C. (2018). Idealized models, holistic distortions, and universality. Synthese, 195(6): 2795-2819.

Rice, C. (2019). Models don't decompose that way: A holistic view of idealized models. The British Journal for the Philosophy of Science, 70(1): 179-208.

Rice, C., Rohwer, Y. & Ariew, A. (2018). Explanatory schema and the process of model building. Synthese, published online.

Richter, C. (1958). Elementary Seismology. San Francisco: W. H. Freemen & Co.

Roca-Royes, S. (2017). Similarity and Possibility: An Epistemology of de re Possibility for Concrete Entities. In B. Fischer & F. Leon (eds) "Modal Epistemology After Rationalism" (pp. 221-46). Cham: Springer

Rohwer, Y. & Rice, C. (2016). How are models and explanations related. Erkenntnis 81(5): 1127-1148.

Salmon, W. (1989). Four decades of scientific explanation. In P. Kitcher & W. Salmon W (eds.) "Scientific explanation" (pp. 3-219). Minneapolis: University of Minnesota Press.

Salmon, W. (1998). Causality and Explanation. New York: Oxford University Press.

Schorlemmer, D., Wiemer, S. & Wyss, M. (2005). Variations in earthquake-size distribution across different stress regimes. Nature, 437(7058): 539-42.

Sheredos, B. (2016). Re-reconciling the epistemic and ontic views of explanation (or, why the ontic view cannot support norms of generality). Erkenntnis, 81(5), 919-949.

Sornette, D. (2006). Critical Phenomena in the Natural Sciences. Berlin: Springer.

Sosa, E. (1980). The raft and the pyramid: Coherence versus foundations in the theory of knowledge. Midwest Studies in Philosophy, 5(1): 3-26.

Sterret, S. (2006). Models of machines and models of phenomena. International Studies in the Philosophy of Science, 20(1): 69-80.

Strevens, M. (2004). The causal and unification approaches to explanation unified-causally. Nous 38(1): 154-176.

Strevens, M. (2008). Depth: An account of scientific explanation. Cambridge, MA: Harvard University Press.

Strevens, M. (2013). No understanding without explanation. Studies in History and Philosophy of Science, 44: 510-515.

Suarez, M. (1999). Theories, Models, and Representations. In L. Magnani, N. Nersessian and P. Thagard (eds.) “Model- Based Reasoning in Scientific Discovery” (pp. 75-83). New York: Kluwer

Suarez, M. (2004). An Inferential Conception of Scientific Representation. Philosophy of Science, 71: 767–779.

Suarez, M. (2015). Deflationary representation, inference, and practice. Studies in History and Philosophy of Science, 49: 36-47.

Suárez, M. (Ed.). (2008). Fictions in science: Philosophical essays on modeling and idealization. Routledge.

Suárez, M. & Cartwright, N. (2008). Theories: Tools versus models. Studies in History and Philosophy of Modern Physics, 39: 62-81.

Svetlova, E. (2013). De-idealization by commentary: the case of financial valuation models. Synthese, 190: 321-37.

Swoyer, C. (1991). Structural representation and surrogative reasoning. Synthese, 87(3): 449-508.

Teller, P. (2001). Twilight of the perfect model model. Erkenntnis, 55(3): 393-415.

Toffoli, T. & Margolus, N. (1990). Invertible Cellular Automata: A review. Physica D, 45: 229–253.

Trout, J. (2002). Scientific explanation and the sense of understanding. Philosophy of Science, 69:212-33.

Turcotte, D. L. (1997). Fractals and chaos in geology and geophysics. Cambridge university press.

Vaidya, A. (2017). "The Epistemology of Modality", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/modality-epistemology/>.

van Eck, D. (2015). Reconciling Ontic and Epistemic Constraints on Mechanistic Explanation, Epistemically. Axiomathes, 25(1), 5-22.

van Fraassen, B. (1980). The scientific image. Oxford: Oxford University Press.

van Fraassen, B. (1994). Interpretation of science; science as interpretation. In J. Hilgevoord “Physics and our view of the world” (pp. 169-187). Cambridge: Cambridge University Press.

van Fraassen, B. (2008). Scientific Representation: Paradoxes of Perspective. Oxford: Oxford University Press.

von Neumann, J. (1955). Method in the physical sciences. Collected Works, 6: 491-498.

Watkins, N., Pruessner, G., Chapman, S., Crosby, N. & Jensen, H. (2016). 25 Years of Self-organized Criticality: Concepts and Controversies. Space Science Reviews, 198: 3-44.

Weber, E., Van Bouwel, J., & De Vreese, L. (2013). Scientific explanation. Dordrecht: Springer.

Weisberg, M. (2007). Who is a modeler? The British Journal for the Philosophy of Science, 58(2): 207-233.

Weisberg, M. (2013). Simulation and Similarity: Using Models to Understand the World. New York: Oxford University Press.

Wilkenfeld, D. (2017). MUDdy understanding. Synthese, 194: 1273-1293.

Wolfram, S. (2002). A New Kind of Science. Champaign, IL: Wolfram Media.

Woodward, J. (2003). Making Things Happen: A Theory of Causal Explanation. Oxford: Oxford University Press.

Woodward, J. (2019). "Scientific Explanation", The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2019/entries/scientific-explanation/>.

Wright, C. (2012). Mechanistic explanation without the ontic conception. European Journal for Philosophy of Science, 2(3), 375-394.

Xia, J., Gould, H. Klein, W. & Rundle, J. (2005). Simulation of the Burridge-Knopoff Model of Earthquakes with Variable Range Stress Transfer. Physical Review Letters, 95: 248501(1-4).

Ylikoski, P. & Kuorikoski, J. (2010). Dissecting explanatory power. Philosophical Studies, 148: 201-19.