Functional Robustness:
A New Framework for Multiple Realization and its Epistemic Consequences

by

Worth Howard Boone III

B.A. in Philosophy, Lewis & Clark College, 2007

M.A. in Philosophy, Simon Fraser University, 2010

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Worth Howard Boone III

It was defended on

October 3, 2019

and approved by

Robert Batterman, Distinguished Professor of Philosophy

Mazviita Chirimuuta, Associate Professor of History and Philosophy of Science

James Woodward, Distinguished Professor of History and Philosophy of Science

Dissertation Director: Edouard Machery, Distinguished Professor of History and Philosophy of Science

Functional Robustness:
A New Framework for Multiple Realization and its Epistemic Consequences

Worth Howard Boone III, PhD

University of Pittsburgh, 2019

In this dissertation, I provide a novel account of multiple realization. My account reframes the concept in terms of causal theories of explanation, in contrast to the original framing in terms of the deductive-nomological theory of explanation. I align my account of multiple realization with the phenomenon of functional robustness, particularly by examining a number of cases of robustness in neural systems. I then explore the epistemic consequences of functional robustness. In particular, I argue that systems that exhibit robustness will tend to violate causal faithfulness, thus posing challenges to causal hypothesis testing and causal discovery. I then consider the proposal that robustness undermines modularity - i.e. the ability of causal relationships within a system to be disrupted independently. I argue that it does not and instead that robustness often is due to feedback control driving systems toward particular outcomes. As a result, robustness will attend failures of acyclicity, not failures of modularity. I conclude by contrasting these epistemic consequences of functional robustness with those traditionally associated with multiple realization.

Table of Contents

List of Tables

List of Figures

Preface

Much individual toil and torment goes into writing a dissertation. That said, it would be hard to oversell the importance that support from my colleagues and loved ones played in getting me through the process.

Intellectually, I have benefited from the work of too many philosophers and cognitive scientists to possibly list here, though I will make some effort. And I have been incredibly fortunate in the support I have received along the way from my family, friends, and significant other; they are truly the ones who have made this possible and deserve the most recognition.

My undergraduate advisors, Becko Copenhaver and Jay Odenbaugh, encouraged me to pursue graduate school and supported my doing so. Much of what I learned from them, regarding philosophy of mind (from Becko) and philosophy of science (from Jay), shaped my interests in these topics and is still evident many years later in this dissertation.

My master's advisor, Kathleen Akins, set me toward the study of the mind-brain sciences. Though my dissertation topic strayed from the perspective on perception and consciousness I gleaned from her, my general interest in the topic endures and it continues to guide other areas of my research. Kathleen remains to this day one of the most inspiring minds I've ever encountered. Just as I likely would not have pursued an MA without Becko and Jay's mentorship, I likely would not have pursued a PhD without Kathleen's.

Of course, my dissertation also owes an immense debt to my committee members. Whatever intellectual value there is in this project owes most to them; whatever faults remain my own. A conversation I had with Mazviita Chirimuuta as a third-year student steered me toward writing on

this topic. She has provided valuable feedback and critique along the way. The influence of Jim Woodward's work on causation and causal explanation is easy to see throughout the dissertation. Jim helped shape the project through several conversations, detailed feedback on drafts, and by sending me his own thoughts and works in progress on related topics. Bob Batterman has been, and continues to be, a model of how to be an excellent academic mentor. My project strayed a bit from the initial proposal in my prospectus, which was more in Bob's wheelhouse. But his influence is no less present. Finally, Edouard Machery has been, in many ways, everything a dissertation advisor should be. He was patient when I needed patience, and impatient when I needed to get moving. He has been an excellent mentor and deserves a great amount of credit for any merits I have as an academic.

Other members of the broader Pittsburgh intellectual community also deserve thanks and recognition. In particular, Sandy Mitchell, whose work I engage with directly in chapter four, was influential in making issues related to robustness prominent during my time at Pitt. Sandy co-organized a conference on Robustness in Neurological Systems in 2015 that was incredibly fruitful both in terms of fodder and feedback for this project. Gualtiero Piccinini, a past HPS grad, has been an invaluable mentor, friend, and coauthor, since we met at a conference and he graciously asked me to be involved in a few projects. Though Wayne Wu was not involved in the development of this dissertation, he was also an excellent mentor along the way and was influential in shaping other of my projects in philosophy of perception. Other faculty members who provided valuable feedback and guidance include Mark Wilson, Jim Bogen, and David Danks. Katie Labuda also deserves special mention for keeping me on top of paperwork and deadlines.

Institutionally, during my graduate education at Pitt, I was fortunate to have access to and support from the Center for the Neural Basis of Cognition, which provided generous traveling

funding and interdisciplinary opportunities, and the Center for Philosophy of Science, where I worked as an administrative assistant from 2015-2016.

I was also fortunate to be part of the graduate student community in the philosophy and HPS departments at Pitt. My cohort, Jeff Sykora and Lei Jiang, later joined by Taku Iwatsuki, supported each other through our comprehensive exams. Mikio Akagi and Joe McCaffrey were a significant part of my intellectual experience at Pitt, both as friends and as members of the PoCs reading group. I would also like to thank Katie Tabb, Tom Pashby, Michael Miller, Marina Baldissera Pacchetti, Haixin Dang, Lauren Ross, Zina Ward, and Siska De Baerdemaeker for their friendship and support throughout my time at Pitt. Morgan Thompson and Michael Mahoney deserve special mention for being there for me when I needed them, always ready with board games and whisky. Julia Bursten and her late dog, Linus Pawling, also were and continue to be wonderful friends. Mary Thibadeau was also a great friend, who alongside Joe McCaffrey, supported me through difficult times.

My dog of 13 years, Sealy Face Dogbody, saw me through the last half of my undergraduate education, my master's degree, and the majority of my PhD. She lived with me through three major moves—two across the country, one to a different country, to three different states and one province. She was my best friend and an amazing adventure copilot along the way.

My current me cadre of critters, Newton, Maize, Moose, and Duck, saw me through the trying days of late dissertation writing with plenty of snuggles and way too much fetching.

My partner, Ashley Jardina, was an incredible support, especially through the final months of dissertation writing. From long nights working by my side to emotional support when I had crises of confidence and felt resigned to not being able to finish, she helped in more ways than I could possibly list here. This dissertation would never have seen the light of day without her love and

support. More than any of that, I appreciate her being a patient, understanding, and kind person who inspires the best in me.

Finally, I would like to recognize and thank my family for their support during my many years of graduate study. My sister, Leighanne, and brother-in-law, Scott, have always been there when I've needed them. My nieces, Anna and Molly, have provided much needed humor and reminders of what it's like to be a curious mind figuring out how to engage with the world. And my parents, to whom I would like to dedicate this dissertation, provided the pillars for my education and supported me unceasingly along the way. My mother, Sherrie, stayed actively involved in my education while I was growing up. She always made sure I had access to and made the most of the best educational opportunities available. My father, Worth, instilled in me at an early age a lasting intellectual curiosity and a desire to find and take joy in the patterns of our world.

1.0 Dissertation Introduction

Brains exhibit remarkable capacities to maintain functions despite substantial variation in the component parts and processes that support those functions. This robustness of neural functions can be found at all levels of organization within the brain. For example, individual neurons show stable electrophysiological properties despite variation in the ion channels that determine those properties. Neural circuits produce stable outputs despite variation in the synaptic strengths between and intrinsic activity of the cells that make up those circuits. And neuroplasticity can enable recovery of function from macroscale damage to entire cortical areas. These different forms of neural robustness are imminently relevant to anyone interested in understanding the mind-brain relation, explanation in neuroscience, and the relationships between different levels of organization in complex systems.

Philosophical debates about the mind-brain relation have, however, failed to make substantial contact with this phenomenon of robustness. This is particularly puzzling given that the concept of multiple realization has been central to these debates since the 1970s. In broad terms, multiple realization is the claim that higher-level properties correspond to a number of distinct lower-level properties. And it is typically cited as a crucial premise in arguments against reductionism and in arguments looking to secure the autonomy of the so-called special sciences. Robustness, at least on its face, would seem to be of patent relevance to multiple realization, as it demonstrates a clear case in which there is stability at the level of the function performed, despite variation in the causal structures that support performance of that function.

Philosophical accounts of multiple realization have, however, had a blindspot to the types of cases robustness presents. Particularly in the context of the mind-brain sciences, these accounts

have tended to focus on the possibility of the same mental state arising in different organisms (e.g. animal pain vs. octopus pain) or in silica (i.e. the possibility of artificial intelligence), whereas robustness points toward a sort of causal heterogeneity underlying stable functions within a particular species or even a particular organism. Some reasons for this have to do with historical coincidence of the scientific state of the art at the time that early debates about multiple realization were taking place. For instance, advances in computer science teased the development of artificial intelligence that might bear similarities to human intelligence. And little was understood about complexity underlying functions within particular neural systems, supporting the assumption that a mental state, like pain, may not be multiply realized within a particular species. This meant looking to computers or other organisms for potential sources of multiple realization.

Another factor influencing where philosophers have been looking for instances of multiple realization involves the background views in philosophy of science that framed its initial discussions. Specifically, those initial discussions were situated within logical positivist views of explanation and reduction. Loosely, according to these positivist views, explanation of some phenomenon consists in logically deriving it from natural laws, and reduction of a higher-level science to a lower-level science consists in showing how the laws of the higher-level science can be logically derived from the laws of the lower-level science. Within this framework, multiple realization is cast as a thesis about the natural kind terms that figure into these natural laws. As a result, the search for examples of multiple realization has largely consisted in finding an example of some mental natural kind, say pain, and making the case that it exists in a number of different systems.

Over the past several decades, consensus in philosophy of science has shifted away from these positivist accounts toward accounts that take causation to be central to both explanation and

reduction. Rather than construing explanation as a matter of derivation from, or subsumption under, natural laws, causal accounts take explanation to consist in illuminating the causal structures that give rise to a phenomenon. Such accounts have been developed with particular focus on higher-level sciences, like biology, cognitive science, social science, etc. This represents a substantial shift in the backdrop that frames debates about multiple realization. Rather than a thesis about natural kind terms that occur in different sciences, multiple realization instead becomes a thesis about causal heterogeneity at a lower level, despite causal stability at a higher level.

This difference in framing brings with it differences in the kinds of phenomena that will be sought as instances of multiple realization. Rather than looking for the same mental state arising in different organisms or in artificial systems, causal frameworks encourage looking for causal complexity underlying causal stability. This invites contact with the forms of robustness mentioned at the outset, as robustness points toward a sort of causal heterogeneity underlying stable functions within a particular species or even a particular organism.

The difference in framing also attends differences in the epistemic significance of multiple realization. As alluded to above, debates about multiple realization have generally focused on its ability to secure the autonomy of psychology (or higher-level sciences, more generally) from neuroscience (or lower-level sciences, more generally). Within the positivist framework, this is the natural way to characterize the epistemic significance of multiple realization. In that framework, multiple realization is cast as a thesis about the natural kind terms that figure into natural laws. If the natural kinds of a higher-level science are multiply realized by the kinds in the lower-level science, then the bridge principles that map between those kinds and are necessary for logical derivation are blocked. The bridge principles are blocked because multiple realization implies that higher-level kinds correspond to heterogenous disjunctions of kinds in the lower-level science.

Such heterogeneous disjunctions are then taken as unsuitable candidates for the nomic bridge principles required for logical derivation. By contrast, in causal frameworks, rather than this comparatively thin epistemic thesis about autonomy, multiple realization instead implies a range of more nuanced epistemic consequences about causal discovery, the structure of causal explanation, how we proceed with causal investigation, and causal hypothesis testing.

My central aim in this dissertation is to articulate this story in greater detail, taking stock of the shift mentioned above in the background views in philosophy of science, demonstrating its relevance to debates about multiple realization, and looking to examples of functional robustness to both substantiate this new account of multiple realization and to draw out its more nuanced epistemic consequences. I proceed as follows.

In chapter 2.0, I consider the shift from positivist to causal models of explanation in more detail and offer a reframed analysis of multiple realization in causal explanatory frameworks. As intimated above, multiple realization has traditionally been cast as a thesis about the relation between kinds posited by the taxonomic systems of different sciences. I show explicitly the ways in which this traditional framing is tied to positivist models of explanation and reduction. I then develop an alternate framing based on causal explanatory frameworks that, in broad terms, characterizes multiple realization as causal stability at a higher level despite causal heterogeneity at a lower level. This framing enables the connections between multiple realization and the notion of functional robustness discussed above. I examine cases of robustness from systems neuroscience that demonstrate this connection, and I show how traditional debates fail to track important features of these cases.

In chapter 3.0, I argue that systems that exhibit functional robustness pose a particular challenge to the problem of causal discovery—i.e. the problem of inferring causal structure from

patterns of probabilistic dependence. Specifically, robust systems are prone to generate failures of causal faithfulness. Causal faithfulness is a condition that grounds many, but not all, causal discovery algorithms in the context of graphical causal modeling. The condition states that any two variables in a system that are causally related are also probabilistically dependent. In systems that exhibit functional robustness, the function in question will often be probabilistically independent of causal variables relevant to performance of that function, thus generating failures of causal faithfulness. I demonstrate such a failure of faithfulness with an example of functional robustness in single neurons. I then discuss the significance of failures of faithfulness for causal inference in neuroscience and in complex systems more generally.

In chapter 4.0, I consider the proposal that robustness undermines the notion of modularity in interventionist theories of causation. Modularity, in general terms, refers to the assumption that components of a causal system make isolated causal contributions to their respective effects. In other words, a system is modular to the extent that any particular causal relationship can be disrupted without altering the other causal relationships within the system. This concept is a core feature of attempts to analyze causation and causal inference in terms of difference-making, particularly interventionist theories of causality (Woodward 2003). Mitchell (2008, 2009) has argued that robustness shows that modularity typically does not hold for biological systems, in particular genetic networks. In this chapter, I argue that Mitchell mislocates the challenge posed by robustness to theories of causal explanation. Rather than failures of modularity, I argue that robustness often indicates cyclic causal structures – therefore indicating failures of acyclicity, not failures of modularity. Cyclic causal structures pose their own challenges to causal inference. I explore those challenges and the resources available to overcome them.

In chapter 5.0, I conclude by contrasting the traditional epistemic consequences associated with multiple realization with those that follow from functional robustness (or causal explanatory multiple realization). Rather than entailing autonomy between different sciences, functional robustness entails a range of consequences for causal inference and explanation. In addition, I highlight several aspects of my discussion that point toward promising avenues for future research.

2.0 Multiple Realization and Robustness

2.1 Introduction

Traditionally, multiple realization (MR) has been understood as a thesis about the relation between kinds posited by taxonomic systems in different sciences (e.g. psychology and neuroscience). This characterization of MR has been heavily influenced by positivist models of explanation, reduction, and the unity of science (Hempel 1942, Nagel 1961, Oppenheim and Putnam 1958), against which early arguments concerning MR (Putnam 1967, 1975; Fodor 1974) were targeted. In this chapter, I explicitly reframe MR in terms of causal explanatory frameworks that better capture explanatory practice in the special sciences. Within such frameworks, MR becomes more a thesis about causal structure than about mapping relations between kinds. This shift in framing exposes connections between MR and the notion of functional robustness in biology and neuroscience.

I proceed as follows. In §I, I show how the traditional framing of MR is tied to outmoded positivist conceptions of explanation and reduction. I then offer an analysis of MR that operates within frameworks of causal explanation that better capture explanatory practice in psychology and neuroscience. In §II, I draw connections between this conception of MR and the phenomenon of functional robustness in biology and neuroscience. I examine in detail two cases of robust functions in neural systems. In §III, I further develop my account by considering and responding to the objection that the account I offer still essentially construes MR to be a relation between kinds. I conclude with brief remarks on the ways this reframing of MR alters the landscape of

debate surrounding nonreductive accounts of the mind-brain relation, and the special sciences more generally.

2.2 Multiple Realization and Causal Explanation

The preoccupation with kinds in philosophical discussions of multiple realization has been, in large part, a holdover from now defunct positivist conceptions of explanation and reduction. The deductive-nomological (D-N) model of explanation, of which the Nagelian model of reduction is an extension, maintains that to explain a phenomenon is to subsume it under some law-like regularity (Hempel 1942, Hempel and Oppenheim 1948, Nagel 1961). This conception of explanation thus assigns the explanatory value of theories to their laws, and by fiat to the (natural) kind terms that figure into those laws. Fodor's (1974) seminal argument from multiple realization to the autonomy of the special sciences targeted the Nagelian model of reduction.[1] As a result, Fodor couched MR as a relation between higher- and lower-level kinds, precluding the formation of nomic bridge principles between higher- and lower-level sciences. This general framing has shaped much of the subsequent debate surrounding MR.

The D-N model, however, has proven to be an inadequate account of explanation in the special sciences, particularly psychology and neuroscience (as well as biology, more generally). A primary reason for this is that neither psychology nor neuroscience deals in laws in the traditional sense (qua universal generalizations), and relatedly, explanations in neither science proceed by subsuming phenomena under regularities (see, e.g., Cummins 1983, Ch1; Craver 2007). To the

[1] Despite the fact that Fodor's title suggests that his target is Putnam and Oppenheim's account of the unity of science. See Shapiro and Polger (2012) for detailed discussion.

contrary, regularities in both psychology and neuroscience provide the targets of explanations, the explananda, rather than the explanantia (Cummins 1983, 2000).

For instance, the "cocktail party effect" denotes a regularity according to which people are able to single out the sound of their names in a noisy environment. Simply citing this effect does little to explain a particular instance of this phenomenon—to do so would be more akin to explaining the sedative properties of opium by appealing to its "dormitive virtue", as famously quipped by Moliere in 1665. Rather, the cocktail party effect characterizes an explanandum, and psychology seeks explanations for why this regularity holds. Similarly, gradually depolarizing a neuron to a membrane potential around -40mV is regularly followed by a rapid depolarization of the cell—the rising phase of the action potential. But again, simply citing this regularity does nothing to explain a particular instance of neural depolarization. Rather, the regularity is the target of explanation into the mechanisms of the action potential. A primary function of taxonomic systems is to capture these sorts of regularities within different scientific domains. That is, with respect to explanatory practice in the special sciences, taxonomic systems and the kinds they posit serve more to characterize explananda than to provide explanantia.[2]

The models of explanation in philosophy of science that have supplanted the positivist framework take causation, rather than subsumption under laws, as the central feature of scientific explanation (e.g. Bechtel 2008, Craver 2007, Salmon 1984, Woodward 2003). While varied in their particulars, what is common to these models is the idea that to explain a phenomenon is to situate it within a causal nexus. Such models fare better at capturing explanatory practice in both

[2] Of course, taxonomic systems also play crucial roles in explanatory practice, but their explanatory value does not consist in capturing nomic regularities or "carving nature at its joints". Rather, the explanatory value of taxonomic systems consists in providing the terms for capturing causal relations between higher- and lower-level analyses. As such, it is those causal relations that do the explanatory heavy lifting in the special sciences, not the kind terms themselves. I revisit this point in more detail in §III, but for now this fast and somewhat loose discussion will do

psychology and neuroscience. For instance, psychologists look to explain the cocktail party effect by analyzing it in terms of functional subprocesses—e.g. selective auditory attention and speech channel separation. Similarly, neuroscientists have explained the rising phase of the action potential by investigating the workings of voltage-gated Na+ channels. In explaining how regularities arise, both psychology and neuroscience look to the causal processes that give rise to these sorts of regularities.

For the most part, causal models of explanation stress decomposition of a system in order to explain how it operates to give rise to some phenomenon. The mechanistic framework (Bechtel and Richardson 1993, Bechtel 2008, Machamer et al. 2000, Craver 2007) currently provides a dominant framework of explanation via decomposition. According to this framework, roughly, to explain a phenomenon is to decompose it into some set of entities and activities that, appropriately organized, explain how the phenomenon was produced (see *Figure 1, b*). Such decompositional explanations, however, only provide half of the story, especially if one is interested in interpolating MR into this framework. The other half consists in upward-directed analyses that explain what a system or phenomenon does within some containing system (see *Figure 1, a*). Such analyses are closely related to what Craver (2001, 2011) calls "contextual explanations" and the explanatory strategy is similar to Bechtel's (2008) notion "reconstituting a phenomenon."

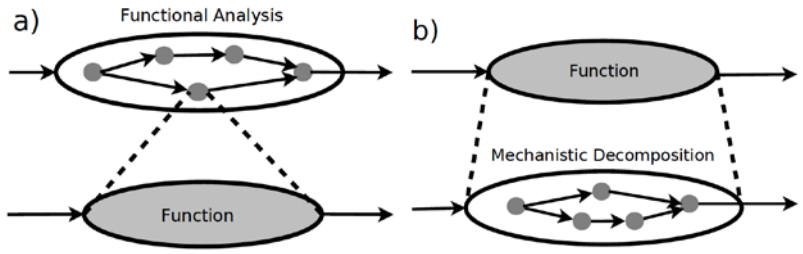


Figure 1: Diagram of the contrast between (a) upward-directed functional analyses and (b) decompositional mechanistic analyses.

Cummins's (1975) account of functional analysis remains one of the canonical ways of capturing this explanatory strategy in both precise and general terms. Cummins's account maintains that functions are ascribed by situating a capacity of a system within an analysis of a capacity of some containing system. In other words, functions are attributed relative to the role they play in analysis of other capacities.

> x functions as a ϕ in s (or: the function of x in s is to ϕ) relative to an analytical account A of s's capacity to ψ just in case x is capable of ϕ--ing in s and A appropriately and adequately accounts for s's capacity to ψ by, in part, appealing to the capacity of x to ϕ in s. (Cummins, 1975: 762)

Both decompositional analyses and upward-directed analyses are crucial to understanding functions within a framework of causal explanation. Upward-directed analyses justify functional attributions, explaining what a system does within some containing system, while decompositional analyses explain how that function is performed by various subsystems. Of course, there is a sort of symmetry between both forms of analysis. A functional analysis can constitute a mechanistic analysis of the function of the containing system, and a mechanistic analysis of a particular function can constitute a functional analysis relative to which the functions of the components of the mechanism are attributed (Piccinini and Craver 2011). Nonetheless, it is useful to keep in mind the distinction between these two forms of analysis in order to interpolate MR into causal-explanatory frameworks.

Juxtaposing functional and mechanistic analyses, MR can be defined as sameness or stability of function (qua causal role in some functional analysis) despite difference in the mechanisms performing that function. Mechanisms are individuated in terms of causal relevance. Two mechanisms are distinct just in case they consist in distinct sets of entities and activities that make relevantly different contributions toward explaining the target phenomenon (in this case, some

function or capacity).[3] Causal relevance can be understood in terms of manipulation and control (Woodward 2003; Craver 2007, Ch3). Thus, a feature of a mechanism is causally relevant if manipulating it while holding the other features of the mechanism fixed alters the phenomenon the mechanism is invoked to explain.

Functional sameness or stability can also be cashed out in terms of causal relevance. Here it is the absence of causally relevant differences (relative to the functional analysis of some containing system) that denotes functional stability. That is, a function is stable across multiple instances just in case whatever differences obtain across those instances are not causally relevant to the role of that function within its containing system. Thus, MR in this framework amounts to the thesis that there are multiple relevantly different causal pathways that converge on a relevantly stable function.[4] For clarity, this thesis of Causal Explanatory MR can be stated as the joint satisfaction of the following two conditions.

Causal Explanatory Multiple Realization (CEMR)

> (1) Two mechanisms are different realizations of a function just in case there are differences between them that would make a difference to performance of the function they explain under controlled intervention (holding all other aspects of the mechanism fixed).
>
> (2) Two instances of a function are relevantly similar just in case whatever differences obtain between them do not make a difference to their roles in

[3] Larry Shapiro (2000, 2004) has also developed an account of MR according to which realizations are distinguished on the basis of causal relevance. As such, Shapiro is a progenitor of the move to frame MR in causal explanatory frameworks. However, Shapiro relies on an intuitive notion of causal relevance, and fails to offer a precise criterion for functionality similarity. Instead, in his earlier work on the subject (though cf. Polger and Shapiro, 2016), Shapiro accepts Kim's (1992) principle of causal individuation of kinds, which in turns leads him to skepticism regarding MR (more on this below). My account thus can be seen as building on Shapiro's work, offering more precise analysis of both causal relevance and functional similarity by tying both to notions of manipulation and control in causal explanation.

[4] By "causal pathway" here I mean a sequence of steps leading from some causal factor to its effect (in this case, some appropriately specified function). This is in rough alignment with the sequential notion of mechanism laid out in Machamer et al. (2000).

> explaining the capacity of a containing system under controlled intervention (holding all other aspects of the functional analysis of that system fixed).

This may seem to invite a puzzle. If two mechanisms are really distinct, then there must be differences between them that are causally relevant to performance of the function in question (by virtue of the conditions I have stipulated for mechanism individuation). So, it would seem that the function cannot be stable across those differences. In other words, it may seem that conditions (1) and (2) are mutually incompatible.

In essence, this apparent incompatibility is the same issue that has motivated a predominant thread of MR skepticism due to Larry Shapiro (2000, 2004). Shapiro has argued that proponents of MR face a dilemma when it comes to distinguishing realizations of a functional kind. The dilemma runs as follows.

> Horn one: if two instances of a functional kind do not differ in a way that is relevant to performance of the function in question, then those instances do not properly correspond to distinct realizations of that kind.

For instance, two waiter's corkscrews that differ only in color are not properly distinct realizations of the kind, corkscrew. It should be clear that the account I've been developing effectively accepts this horn (though I resist framing the issue in terms of kinds—more on this is §III).

> Horn two: if two instances of a functional kind do differ in a way that is relevant to performance of the function in question, then it would seem that there are genuine causal differences between those instances and thus that they correspond to distinct kinds.

The consequent in horn two follows from acceptance of the principle of causal individuation of kinds (Kim 1992). For a mundane example, consider spark-ignition and compression-ignition as relevantly different ways of powering an engine. It would be misleading to say that spark-ignition engines and compression-ignition engines are different realizations of the same kind because the

causally relevant difference between them seems to track a difference in kind rather than a difference in realization of the same kind.

The account of MR developed above offers a way around this dilemma. The advance offered is a precise criterion for functional stability that does not get bogged down in issues of kind individuation—i.e. condition (2) of CEMR. Specifically, stating the issue in terms of kinds invites application of a general criteria of "kindhood" (like Kim's principle of causal individuation) that may not actually be relevant to scientific instances of MR. If we instead interpolate Shapiro's second horn into CEMR, it is plain to see that the issue just points to the puzzle outlined above: if two mechanisms are really distinct, then there are differences between them that are causally relevant to performance of the function in question; and so it would seem that the function cannot be stable across those differences. The issue here is just to see how relevantly different causal pathways can converge on a (relevantly) stable function. And the apparent puzzle can be resolved by noting that there may be differences in other causally relevant features of a mechanism that compensate for some particular causally relevant difference to produce a stable function. In such a case, a particular causally relevant difference is sufficient for distinguishing two realizations of a function, while the compensatory differences among other causal factors in turn enable stability (or relevant similarity) of function. This is in fact commonplace in biological systems, as will become clear in the next section.

2.3 Multiple Realization as Functional Robustness

To this point, I have argued that multiple realization can be understood in causal explanatory frameworks as similarity in what a system does (relative to a functional analysis of some

containing system), in spite of differences in how it does it (specified by some set of mechanistic decompositions). This conception of MR is tied purely to the structure of causal explanations rather than to features (e.g. nomicity, causal individuation, projectibility) of kinds that figure into different taxonomic systems (cf. Fodor 1997, Kim 1999, again more on this in §III). The notion of (functional) robustness maps fairly precisely onto this causal explanatory characterization. The aim of this section is to flesh out this connection and to provide empirical examples of robustness that thereby substantiate this account of MR.

In the first place, there are several related notions of robustness that have received substantive attention in both the biological sciences and philosophy of science and should be distinguished. The first, which I will term "methodological" robustness was introduced by Levins (1966) and has been championed in philosophy of science with the work of Wimsatt (1980, 1981, 2003) and more recently Weisberg (2006) and Schupbach (2018). Robustness in this sense means "accessible (detectable, measurable, derivable, produceable, or the like) in a variety of independent ways" (Wimsatt, 2003). Robustness also has close ties to notions of "stability" and "invariance" that have been cited as criteria on explanatory generalizations that move away from the standard (positivistic) conception of "laws of nature"—qua universal generalizations (e.g. Mitchell 1997, Woodward 2003). The notion of robustness in which I am interested is related to both of these concepts but is nonetheless distinct in relevant ways. This notion, which I call "functional" robustness, is the robustness of some effect produced by a system over variation in or perturbations to the components and properties of that system (Mitchell 2008). This latter notion has been central to recent research in biology and has played a crucial role in systems neuroscience.

Kitano (2004) defines functional robustness as "a property that allows a system to maintain its functions despite external and internal perturbations" (Kitano 2004: 826).[5] The concept is of central relevance to genetic networks in which a large amount of redundancy is built to ensure that systems do not break down in the face of, e.g., minor errors in genetic transcription. Functional robustness is also of central relevance to engineering science in systems in which stable effects must be maintained in response to a range of environmental perturbations. For instance, the autopilot system in modern airplanes is designed to maintain a flight path against a range of changes in atmospheric conditions through compensatory adjustments to various flight mechanisms; similar for cruise control in maintaining a constant speed in automobiles. It is no coincidence that the systems in which functional robustness figures most crucially are also those in which the notion of function has typically been employed—i.e. biological systems and engineered artifactual systems. In such systems there are selective pressures for effects rather than causes, and so there is need for stability in what a system does that supersedes stability, and in fact errs toward variation, in how it does it.

Some initial distinctions are in order before turning to specific examples of robustness in neural systems. Robustness, in the sense discussed by Kitano and other biologists, should be distinguished from the more general concept of functional stability. For any function there is some normal range of variation in its mechanisms over which it may be stable. For instance, spark-ignition engines can combust a range of air-fuel mixture ratios (roughly, between 8:1 and 18:1) that are regulated by a carburetor. The function of the engine is thus stable over this range of ratios. But this form of

[5] Kitano uses the term "biological robustness" because he is interested specifically in how the causal notion of robustness applies to genetic networks. For consistency and to keep clear these multiple senses of the term "robustness," I continue to use my more general term, "functional robustness," in reference to Kitano's work and throughout the remainder of the dissertation.

stability is weaker than or at least distinct from that implied by the concept of robustness of interest to biologists. Functional robustness is a subclass of functional stability that involves some form of reorganization of a system in order to maintain function in the face of perturbations. The concept of reorganization here implies different causal contributions from other components of the mechanism in question.

Here a further distinction can be drawn between (at least) two ways in which reorganization can arise: redundancy and distributed robustness (Wagner 2005). Redundancy occurs when a system maintains function via some redundant mechanism that fills in for a perturbed component. For instance, imagine a spark-ignition engine with a backup carburetor that fills in should the primary carburetor become damaged. In such cases, the redundant part plays the same causal role in the system. As such, functional robustness via redundancy does not qualify as a genuine instance of MR based on the account offered in §I (due to the lack of causally relevant differences in the mechanisms that explain such functions). By contrast, distributed robustness occurs when many different parts play a range of different causal roles that compensate for effects of perturbations. Though there is no easy analog for engines, it might be something like a spark-ignition engine having the capacity to reorganize itself into a compression-ignition engine and sort out a way of converting gasoline into diesel in response to a carburetor failure. It sounds ridiculous in the context of engines, but something like this seems to be remarkably common in certain biological systems.

In systems neuroscience, the study of robustness very much is a science of multiple realization. Neuroscientists concerned with robustness strive to understand many of the features of the mind-brain relation that motivated early work on multiple realization—e.g., the stability of macrolevel regularities to microlevel variation (Putnam 1975; Fodor 1968, 1974), the fact that the same

psychological kinds seem to be realized and realizable in different organisms and artificial systems (Putnam 1975), the fact that psychological functions can be stable over changes that occur in the course of development, and the fact that psychological function can be stable over substantial neural damage (Block and Fodor 1972). Despite the patent relevance of functional robustness to MR, it has received scant attention from philosophers of mind. This is likely because the obviousness of the connection has been obscured by the positivistic hangover that has shaped debates about MR. However, with the causal explanatory framing of MR I offered in §I, it is not much work to connect these two concepts.

To see how MR and distributed robustness relate, it will be helpful to first examine some instances of robustness in neural systems. In long-lived organisms—including humans and lobsters (the purpose of this odd association will become apparent)—individual neurons can persist and function properly for decades. By contrast the proteins and receptors that modulate the electrophysiological properties of those neurons are decaying and being replaced on timescales of minutes to hours and days to weeks. As a result, the features that determine a neuron's electrophysiological properties are in a continuous state of flux. And yet those electrophysiological properties are remarkably stable over time. This poses a mystery regarding how this stability is achieved. As Marder and Goaillard (2006) state the problem,

> [E]ach neuron is constantly rebuilding itself from its constituent proteins, using all of the molecular and biochemical machinery of the cell. This allows for plastic changes in development and learning, but also poses the problem of how stable neuronal function is maintained as individual neurons are continuously replacing the proteins that give them their characteristic electrophysiological signatures. (Marder and Goaillard 2006: 563)

The electrophysiological signatures here refer to both the response properties of neurons as well as their intrinsic excitability. These features are determined by proteins and receptors that enable and modulate the flow of ions across the cell membrane. Experimental work has revealed that

individual neurons exhibit many-fold variability in their expression of particular ion channels (see Marder and Goaillard 2006, and Marder 2011 for review). Despite this variability in channel density (i.e. ion channels per surface area), those same neurons exhibit remarkably similar electrophysiological profiles.

This presents a puzzle. The influx and outflow of ions is what explains the characteristic fluctuations in membrane potential that constitute the electrophysiological properties of a given neuron. So how is it that the channel densities that determine the rates of the influx and outflow of ions can vary while the electrophysiological properties remain stable?

Note that the puzzle encountered here is the same puzzle posed at the end of §I. The problem there was to understand how features that are causally relevant to the performance of a function can vary while that function remains stable. The answer I alluded to was that there can be compensating differences in other causally relevant features that explain this stability. And indeed, computational models demonstrate that a variety of combinations of ion channel densities can give rise to similar electrophysiological profiles in model neurons. These results show that very different combinations of channel densities can produce the same intrinsic bursting profiles (Golowasch et al. 2002, Prinz et al. 2003). Taken together with the observed variability in channel density, it can be inferred that neuron electrophysiology is tightly regulated by compensatory mechanisms to maintain target levels of activity. And indeed, the existence of such compensation has been confirmed in genetic knockout experiments (Guo et al. 2005, Nerbonne et al. 2008).

What this all suggests is that the functions of individual neurons provide an instance of the sort of MR outlined in §I. That is, neurons often exhibit stable functions despite variation in the mechanisms that allow and explain performance of those functions. Again, mechanisms are individuated on the basis of features that are causally relevant to performance of a function, and

functions are specified relative to a functional analysis of some containing system. The densities of ion channels are the primary component features that determine a neuron's electrophysiological profile. And so different combinations of ion channel density distinguish different mechanisms that explain the electrophysiology of a given neuron. It is generally taken for granted in neuroscience that the functions of neurons are determined by their electrophysiological profiles. However, to bring those functions into alignment with the causal explanatory framework from §I, they must be specified relative to some functional analysis of a containing system.

Rarely, and usually only in very simple organisms, do the activities of individual neurons figure directly into explanations of an organism's behavior. To understand how the activities of individual neurons contribute to the behaviors of whole organisms, it is often necessary to first determine the roles those neurons play in intermediate-level causal structures. Specifically, the functions of individual neurons are most often specified relative to their roles in ensembles of neurons—circuits and networks. It is then the functions of these circuits and networks that figure into explanations of simple behaviors. (We do not currently have well-articulated explanations for more complex behaviors in large part because there are likely to be more tiers of intermediate-level causal structure of which we currently have impoverished understanding.) So, in order to gain insight into how circuits operate and what functions they perform, neuroscientists look to simpler systems.

The stomatogastric ganglion (STG) of decapod crustaceans is a small network of about 30 neurons in the stomatogastric nervous system that generates and maintains various motions involved in digestion. There are two main functional networks in the STG: the pyloric network and the gastric network. Both networks produce patterned motor outputs that control particular aspects of crustacean digestion. The primary function of the pyloric network is to generate a three-

phase motor pattern that traffics food particles through the pylorus in a wave of peristaltic motion. This triphasic rhythm has received extensive attention from systems neuroscientists looking to understand the ways in which activities of individual neurons combine to produce characteristic circuit function. Analysis of the triphasic rhythm has shown that the inference from the functional roles of individual neurons to the functions of neural ensembles is far from trivial. Here again, MR is rife in the structure of these interlevel causal explanations.

Prinz et al. (2004) demonstrated that the pyloric rhythm can be generated by vastly different values of the parameters that define the pyloric circuit. Using a simplified (three neuron) model of the pyloric network, they created a database of all possible combinations of synaptic strength and intrinsic electrophysiological properties of the cells in the circuit. Out of more than 20 million possible combinations of circuit parameters, more than 4 million sets of those parameters generated rhythms that exhibited the characteristic three-phase signature of the pyloric rhythm—call this the broad criterion. And of those, 11% (just under half a million sets of parameters) satisfied narrowly defined biological criteria derived from in vitro recordings of pyloric rhythms from a large sample of lobster preparations—the narrow criterion. Importantly, the parameter sets that satisfied both the narrow and broad characterizations of the pyloric rhythm included all possible parameter values for both the intrinsic properties of the individual neurons as well as almost all possible values (with one variable having a restricted range) for the synaptic weights between the neurons in the circuit. So, no particular component or connection within the circuit dominates circuit function.

Given Prinz et al.'s data, the pyloric rhythm provides a clear instance of multiple realizability of the sort outlined in §I. The intrinsic properties of the neurons comprising the pyloric network and the synaptic weights within the network are precisely the features that are causally relevant to

production of the triphasic rhythm. So, the different mechanisms that explain the triphasic rhythm correspond to the different sets of parameter values (i.e. particular network configurations) that support the function. It is these specific network configurations that explain how the pyloric rhythm is generated in any particular case. But there is no universal answer to this "how" question. That is, there is no single mechanism that is responsible for production of the pyloric rhythm. Just as in the case of electrophysiology of single neurons, tuning of other causally relevant features of the network (other synaptic weights and intrinsic properties of component neurons) allows multiple sets of parameter values to converge on a stable target output (e.g., see *Figure 2*).

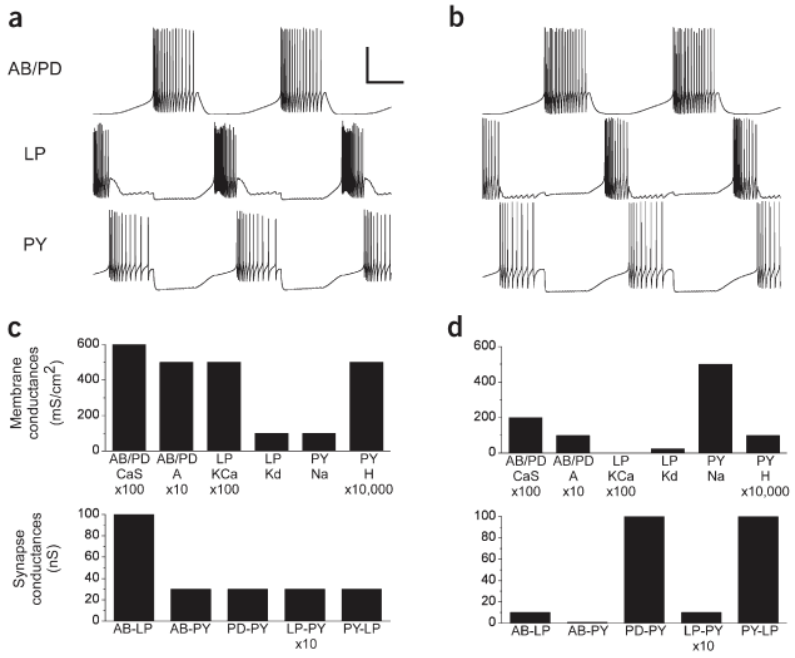


Figure 2: Vastly different sets of parameter values (c,d) give rise to nearly identical circuit function (a,b). From Prinz et al. (2004).

That target output—i.e. the function of the pyloric rhythm—can be specified relative to its role in the functional analysis of crustacean digestion. Thus, the rhythm functions to open (and then close) the pylorus and to produce a wave of peristaltic motion to traffic food particles through the pylorus. Prinz et al.'s broad and narrow criteria correspond to two different ways in which the role

of the pyloric network in this digestive capacity can be analyzed. The broad criterion specifies relevant similarity simply in terms of production of a three-phase rhythm. There are empirical reasons for thinking this is a reasonable criterion: specifically the motoneuron that mediates between the pyloric network and the pylorus seems to act as a sort of temporal filter, so the relevant information from the network is just the order and timing of the firing of the neurons in the three-phase sequence. The narrow criterion constrains the function to the range of biological variability of circuit output observed in vitro. Again, there are theoretical motivations—in this case the lack of certainty that order and timing are the only causally relevant features of pyloric network output—to take this as the criterion of relevant functional similarity. These functional analyses specify what the circuit is doing within the organism and as such determine the range of relevant similarity (or acceptable variability) in the output of the circuit.

This can be made more precise by specifically examining the two functional outputs (a, b) and network configurations (c, d) in *Figure 2*. The two functional outputs are not exactly similar, but they are well within the range of observed variability of in vitro recordings of pyloric network output. On the other hand, the two network configurations—i.e. the two mechanisms realizing those functions—are relevantly different. Any controlled intervention changing one of the parameters in configuration (c)—e.g. the KCa conductance (500mS/cm2) of the LP neuron—to its corresponding value in the second configuration (d)—in this case, completely blocking KCa conductance (0mS/cm2) of the LP neuron—would cause catastrophic failure of the network rhythm. It is the tuning in other network parameters—i.e. the other causally relevant features of the mechanism—that enables the two networks to produce relevantly similar functions despite these differences. Thus, the pyloric network provides a clear instance of MR in the sense outlined

in the first section: relevant similarity in function despite relevant difference in the mechanisms that perform that function.

2.4 Kinds Reconsidered

My central goal to this point has been to provide an analysis of MR, substantiated with empirical examples, that moves away from positivist conceptions of explanation and reduction and operates instead within causal explanatory frameworks. In such frameworks, I have argued MR should be construed as a thesis about the structure of causal explanations rather than a thesis about relations between kinds that figure into different taxonomic systems. One might object that MR in my framework is still, fundamentally, a thesis about the relation between higher- and lower-level kinds. That is, my framing merely offers a different analysis of the kinds involved in putative instances of MR, not a complete abandonment of the concept or utility of kinds in this context. In short my response is that while it is certainly possible to interpolate some notion of kinds into this framework, the relevant shift in the ways kinds are characterized negates much of the philosophical debate that has focused on kinds in the context of MR. The aim of this section is both to develop this objection and to spell out my response in more detail.

Recall that the causal explanatory framework outlined in §I consists in two parts: (1) realizations of a function consist in the mechanisms that explain how that function is performed; (2) functions are specified as causal roles within a functional analysis of some containing system. Despite my insistence to the contrary, it would seem there's a natural way to interpolate kinds into this framework. Specifically, the mechanisms that realize functions may be thought to correspond to lower-level kinds, while functions, qua causal roles within some containing system, may be

thought to correspond to functional kinds in much the way traditional accounts of MR have assumed. On this modified framing, MR would still amount to the traditional claim that there is a many-one relation between lower-level kinds (mechanisms) and higher-level kinds (functions). Also, notably a causal explanatory framing of type-identity theory could be then couched in this framework as the claim that there is a one-one relation between mechanisms and functions.[6]

To reiterate the challenge, interpolating kinds into CEMR involves (a) identifying mechanisms as lower-level kinds and (b) identifying the functions specified in functional analysis as higher-level (functional) kinds. I'll respond to each of these claims in turn. With respect to (a), identifying mechanisms as kinds (qua members of some scientific taxonomic system) is more problematic than it may appear at first glance. Consider the mechanism of a generic snap mousetrap. That mechanism consists in something like the following. The mousetrap is set by lifting the hammer off the platform, pulling it back against the force of the spring, placing the holding bar over the hammer/spring, and then engaging (and baiting) the catch that holds the hammer in place. When the catch is released, the potential energy of the spring is converted into kinetic energy causing the hammer to slam down on the other side of the platform. Note that this mechanism is a complex causal process; it is not a kind in anything like the traditional philosophical sense, and it is certainly not a simple element of a taxonomic system in terms of which mousetraps might be analyzed.

A taxonomy of the components of a snap mousetrap might consist in a list of elements like: platform, hammer, spring, holding bar, catch. These are the elements in terms of which the function of the mousetrap may be analyzed. But the mechanism itself is a complex of these taxonomic elements, and it is their arrangement and causal coordination that explains how snap mousetraps perform their functions. From the other direction, note that snap mousetraps could be construed as

[6] I take this to be roughly the view defended, albeit in different terms, by Polger and Shapiro (2012, 2016).

a particular kind in a taxonomy of mousetraps—among others like glue, poison, or electric mousetraps. Generic mapping relations (one-one, one-many, many-one) between this higher-level taxonomic system and the lower-level taxonomic system of mousetrap components do not track anything interesting about explanations of the operations of these different kinds of mousetraps.

This is nothing peculiar to toy examples. The same applies to well-worn scientific examples like the mechanism of the action potential. Action potentials are, plausibly, activity-kinds in cellular-level neuroscientific taxonomy. The subcellular-level taxonomy in terms of which the action potential is explained consists in kinds like: voltage-gated Na+ channels, voltage-gated K+ channels, plasma membranes, and Na+ and K+ ions. The mechanism itself, of course, belongs to neither of these taxonomic systems. Rather, as in the mousetrap example, the mechanism is a complex causal process that here involves activation of voltage-gated Na+ when a neuron's membrane potential depolarizes to some threshold, usually between -55mV and -40mV, causing a rapid influx of Na+ ions, and so on.

This may seem like a nitpicking point, but the general framing of MR as an issue of the alignment of taxonomic systems continues to be the default view for many philosophers (see, e.g., Polger and Shapiro 2016). Now, while this all suggests there are good reasons to resist thinking of mechanisms as kinds in the sense of simple terms in a taxonomic system, one may still object that mechanisms must be kinds because they have scope. That is, mechanisms are not simply token causal processes, but rather are causal process-types that apply across multiple instances. This is borne out, for instance, in both the examples considered above. The mechanism of the snap mousetrap does not just explain how *this particular mousetrap* operates, but rather explains how *snap mousetraps in general* operate. And the same is true mutatis mutandis of the mechanism of the action potential. Presented with such examples it may be tempting to think that mechanisms

are actually a sort of functional kind coextensive with the functions they realize. Indeed, the mechanism of the snap mousetrap described above is a sort of functional description; and everything that satisfies that description is a snap mousetrap, and every snap mousetrap satisfies that description. Again, the same seems to be true mutatis mutandis of the mechanism of the action potential.

But here we have to be careful and thinking in terms of kinds (and generalizing from examples of this form) muddies the waters.[7] Mechanisms and the functions they realize need not be coextensive. To insist that they are would be to rule out causal explanatory MR tout court. The causal explanatory framing of MR outlined in §I distinguishes the individuation conditions of functions from the individuation conditions of mechanisms: recall that condition (1) lays out the individuation conditions of mechanisms, condition (2) the individuation conditions of functions. The coherence of CEMR thus shows the identification of functions and mechanisms to be conceptually problematic, and the cases of robustness from §II show that identification to be empirically problematic. So, it would seem to be a mistake to construe mechanisms as functional kinds, just as above it proved problematic to construe mechanisms as structural kinds in any straightforward sense. Thus, it seems that there is no straightforward way to interpret mechanisms as kinds in any classical sense of the term. And moreover, foisting the concept of kinds onto mechanisms seems to invite confusion regarding the relation between mechanisms and the functions they perform.

[7] This is one diagnosis of a problem with the type-identity theory that Polger and Shapiro (2012, 2016) defend. They generalize from toy examples like corkscrews and scientific examples that involve quite general characterizations of mental/neural functions to reach the conclusion that functions are, in the vast majority of cases, identical with the mechanisms that realize those functions.

We can now take a closer look at (b), the identification of functions specified in functional analysis with higher-level (functional) kinds. On its own, this proposal is not as fraught as (a) but does bear its own pitfalls. The point of maintaining that functions are always attributed relative to a functional analysis of some containing system is to build a significant amount of context-sensitivity into functional attributions. Specifically, functional analyses play the crucial role of determining the relevant grain of generality at which functions are specified.

Take hearts as an example. At a most general level the function of a heart can be specified relative to its role in a circulatory system—viz. pumping nutrient fluids. At such a general level, there is no motivation to distinguish between the functions of insect hearts and vertebrate hearts. That is, any organ embedded in a circulatory system that pumps nutrient fluids functions as a heart in this general sense. However, if we perform more fine-grained functional analysis of circulatory systems, and consider the sorts of nutrients those fluids supply (e.g. oxygen) and the ways those nutrients are supplied to body parts (i.e. through open or closed circulatory systems), insect hearts and vertebrate hearts no longer perform the same function.

At this grain of functional analysis, note that fish hearts and human hearts do perform the same function. However, if we analyze the functions of vertebrate hearts in terms of their role in blood oxygenation, fish hearts and human hearts no longer perform the same function. In fish circulatory systems, the heart simply functions to circulate blood (via a single pass per circuit), with the blood picking up oxygen from the gills en route to the rest of the organs and body parts. In human circulatory systems, the heart serves a dual function (via two passes per circuit) of circulating deoxygenated blood to the lungs and oxygenated blood to the rest of the body.

The point of these examples is that the relevant function the heart performs changes depending on the way its role within its containing system (the circulatory system) is analyzed. In the context

of circulatory systems generally, hearts function to pump nutrient fluids simpliciter; in the context of open circulatory systems, hearts function to pump nutrient fluids (for insects, hemolymph) diffusely throughout the body; in the context of closed circulatory systems, hearts function to pump nutrients and oxygen-transporting red blood cells through a system of blood vessels; and so on. The advantage of tethering functions to functional analyses is that doing so keeps this context in place and encourages clarity regarding the degree of generality at which those functions are specified. Thinking in terms of functional kinds, on the other hand, invites decontextualization of functions ("the heart functions to pump blood"), and encourages lack of clarity with respect to degree of generality.

Further there is a close connection between functional analytic context and the criteria that determine and distinguish between realizations of a given function that risks getting lost when functions are construed as kinds. For instance, are insect hearts genuine realizations of the functional kind, heart, even though they don't "pump *blood*"? What differences between realizations of hearts are causally relevant to their ability to "pump blood"? Are two-chambered hearts and four-chambered hearts two different kinds or different realizations of one kind? These questions are too vague to be determinately answered in the absence of the context provided by some more precise functional analysis of the circulatory systems in which hearts are embedded. Again, the ability to clearly determine and distinguish between realizations gets lost in decontextualized functional attributions—i.e. subsumption under functional "kindhood".

Of course, one could argue that I'm not really giving up the notion of functional kinds, but rather that I am advocating a radical contextualization of functional kinds. After all, functions, even when tightly coupled with functional analyses, do have scope beyond token instances. Thus, although I may be denying that hearts are a univocal functional kind, what I'm actually advocating

is that hearts can be divided into many different functional kinds that correspond to different degrees of generality depending on circulatory system context. There does seem to be something to this. We do distinguish between insect hearts and vertebrate hearts, between fish hearts and mammalian hearts, and these distinctions do seem to track differences in scope, and thus may be construed as tracking differences in kind.

My reply to such a counter is similar to that which arose in the discussion of mechanisms as kinds. On one hand, I can only concede that this sort of stripped-down notion of kinds (qua any predicate with scope) can be applied to my account of functions. On the other hand, I can certainly urge caution in the ways philosophical habits of thought regarding kinds are applied within such an account; and I can further point out that a highly context sensitive notion of functional kinds fails to make solid contact with a significant thread of philosophical discussion regarding MR.

On this latter point, I can offer some more specific remarks. Due to the positivist backdrop of most philosophical debates about MR, the focus on functional kinds has centered on their ability to figure into special science laws (rather than causal explanations). For instance, Fodor (1997) argues that functional kinds are vindicated by their role in special science laws (whereas heterogeneous disjunctions are not appropriately nomic, and so lower-level "laws" that attempt to capture higher-level, multiply realized regularities are not in fact laws). By contrast Kim's (1992) MR skepticism is grounded in the claim that scientific kinds must be individuated on the basis of causal powers, which has ties to Shapiro's (2000, 2004) MR dilemma discussed in §I. And further, Kim argues the hallmark of natural laws is that they are projectible generalizations—i.e. a confirming instance of a lawlike generalization of the form "All Fs and Gs" provides reason to believe that Fs will be Gs in all contexts. Kim argues that generalizations involving multiply realized kinds are not projectible in this way, and so MR ought to be rejected.

The issue for both Fodor and Kim in the context of this debate hinges on what criteria one adopts for nomicity or naturalness of kinds. But such criteria fail to gain traction with a highly contextualized notion of functional kinds. Contra Fodor, the generalizations such kinds figure into do not aim for lawlike status; they are confined to their functional analytic contexts. Similarly, their inductive projectibility is confined to functional analytic context; there are no ambitions to project universally. But based on the account I've been developing, none of this should garner pessimism regarding the prospects of MR. One can either give up the prospects for regarding contextualized functions as kinds, or one can insist on their characterization as kinds and give up direct contact with these traditional ways of framing MR debates. Once we shift the debate into the context of causal explanations, issues regarding nomicity and lawfulness are exposed as red herrings that the philosophical conversation ought to move beyond.

2.5 Conclusion

The aims of this chapter have been largely positive. In the first place, I provided an analysis of MR that moves away from positivist conceptions of explanation and reduction and operates instead within causal explanatory frameworks. In such frameworks, I argued that MR can be construed as a thesis about the structure of causal explanations rather than a thesis about relations between kinds that figure into different taxonomic systems (granted the caveats of §III). My second main aim has been to provide empirical examples that substantiate this notion of MR by drawing connections between MR and functional robustness in systems neuroscience. The traditional philosophical considerations that have surrounded MR (e.g. nomicity, projectibility, causal individuation) fail to adequately track important features of these empirical cases. This should perhaps be unsurprising

given that those debates are based on an outmoded framework of explanation and reduction in the special sciences. One might worry, however, that tailoring an analysis of MR to these cases sacrifices much of the philosophically interesting features of MR. In the chapters to follow I show that quite the opposite is true. The connection between MR and robustness provides a range of important and interesting epistemic consequences for causal inference and causal explanation.

3.0 Robustness and Causal Faithfulness

3.1 Introduction

My argument in chapter 2.0 encouraged a shift away from the traditional framing of multiple realization (MR). Specifically, I argued that MR should be situated within causal models of explanation rather than the deductive-nomological (DN) model of explanation. Philosophical interest in MR has always ultimately been motivated by its epistemic consequences. The DN-model of explanation takes laws and natural kind terms as the fundamental features of science. The traditional take on the epistemic consequences of MR has focused in turn on MR's ability to secure the autonomy of higher-level laws from lower-level laws. Since I encouraged a shift away from the DN-model in the framing of MR, one would expect a corresponding shift in the epistemic implications that follow from my causal explanatory framing of MR (what I termed CEMR). Specifically, rather than supporting the autonomy of laws, CEMR ought to have implications for the formation of causal explanations, the discovery of causal structure, and the testing of causal hypotheses.

In this chapter, I draw out a subset of these implications by exploring the consequences of robustness for the causal faithfulness condition (CFC). Robustness, in the sense that exemplifies CEMR, refers to a system's capacity to maintain functions despite relevant differences in the causal structures that realize those functions. The CFC is a condition that relates information about probabilistic dependences between variables to representations of the causal relationships between those variables. It has important implications for both causal discovery and causal hypothesis

testing. I argue that systems that exhibit robustness are likely to violate the CFC, and I explicate the consequences of such violations of faithfulness for causal inference in those systems.

I proceed as follows. In §I, I introduce the general framework for causal inference using directed acyclic graphs—including the causal Markov condition and the causal faithfulness condition—and I explicate the most common argument cited in defense of the CFC. In §II, I offer an analysis of the general dialectic surrounding the faithfulness condition, including objections and responses. In §III, I examine a case of robustness from cellular neuroscience, show how this case exhibits a clear violation of faithfulness, and show how this example generalizes across many levels of analysis in neuroscience. In §IV, I situate this example with respect to the debates surrounding faithfulness outlined in §II and discuss the significance of violations of faithfulness more generally for causal analysis of complex systems.

3.2 Graphical Causal Modeling and the Causal Faithfulness Condition

Graphical causal models (GCMs) represent the causal structure of a system in terms of *nodes*, representing variables of interest, connected by *edges*, representing causal relations between those variables. Directed acyclic graphs (DAGs) are a form of GCM that contain information about the direction of the causal links between nodes and contain no directed cycles (bidirectional pathways between nodes). In themselves, such models contain no information about quantitative relationships between variables in a given graph—i.e. probability distributions over those variables. That is, graphical models provide only tools for representing causal relationships; they do not directly entail anything about how those causal relationships manifest in quantitative relationships between variables.

Of course, being able to move between probability distributions and causal structure is desirable. Such inferences license *causal hypothesis testing*—testing causal structures by deriving statistical predictions from them and checking them against data—as well as *causal discovery*—inferring causal structure from statistical data. In order to form these inferences some additional principles linking causal structure to statistical data have to be adopted. Two conditions that have been influentially proposed to fill such a role are the Causal Markov Condition and the Causal Faithfulness Condition (Spirtes, Glymour, and Scheines 1993).

It is useful to introduce the Causal Markov Condition (CMC) by characterizing it as a generalized form of the screening off relation in Reichenbach's common cause principle (Reichenbach 1956—for discussion see Cartwright 1999). The common cause principle captures the idea that correlations between events often point *not* to a causal relation between those events, but instead to a common cause of both events. A common cause is said to screen off a correlation between two variables just in case the two variables are uncorrelated (probabilistically independent) when one conditions on the common cause variable. Take for instance the example in *Figure 3.* Psoriasis and depression are positively correlated—i.e. P (psoriasis & depression) > P (psoriasis) * P (depression). However, that positive correlation is screened off when one conditions on alcohol abuse—i.e. P (psoriasis & depression | alcohol abuse) = P (psoriasis | alcohol abuse) * P (depression | alcohol abuse).

Figure 3: The screening off relation.

The CMC generalizes this screening off relation to the joint probability distribution over all variables in a causal graph. Informally, the CMC states that if two variables are correlated, then either they share a common cause or they are causally related.[8] In other words, any correlation that cannot be screened off by a common cause implies causation. In this way, the CMC stipulates a sufficient condition on two variables being causally linked—i.e. it is sufficient that they share a correlation that cannot be screened off. Thus, X is a direct cause of Y if, conditional on all other variables in the causal graph, X and Y are dependent. The formal definition more often takes the form of the contrapositive: all variables in a system that are not causally related are probabilistically independent. In this direction of inference, the CMC stipulates a sufficient condition on variables being probabilistically independent—i.e. it is sufficient that they not share a causal link. In other words, the CMC provides an entire set of independence and conditional independence relations for a given causal graph.

The CMC is appealing because it captures intuitions about how causal structure should entail conditional probabilistic independencies. If two variables are not causally related, then those variables should be probabilistically independent, conditional on their direct causes—or 'parents'—provided that neither variable is a descendant of the other. Formalizing this intuition allows a first step toward mapping between causal models and quantitative relationships between variables. The CMC contains information relevant to both causal hypothesis testing and causal discovery. With respect to causal hypothesis testing, the CMC entails that, conditional on their direct causes, all variables in a causal graph will be probabilistically independent of all nondescendant variables. With respect to causal discovery, the CMC entails that any correlation

[8] A third possibility is that the variables share a common effect (called a *collider*), in which case conditioning on the collider will yield a dependence between the variables that are otherwise conditionally independent.

between two variables that holds despite conditionalization on all other variables in the graph implies that those two variables are causally linked. The direction of that link can also be inferred in that the effect will be probabilistically dependent on the cause, but not vice versa. (See *Table 1* for reference.)

However, by itself the CMC is not sufficient to either generate substantive statistical hypotheses from causal graphs or to derive substantive causal graphs from statistical data. This is because the CMC alone is too easily satisfied—inferring probabilistic relationships from causal structure, it only entails information about what independences (both conditional and unconditional) follow from a causal graph and does not entail anything about the dependences that should follow. From the other direction, inferring causal structure from probabilistic relationships, the CMC only entails information regarding which variables are causally related and does not entail anything regarding which variables are not causally related. Note, for instance, that if all variables in a graph were probabilistically independent, the CMC would be satisfied trivially. Similarly, if every variable in a graph were connected as a direct cause to every other variable in that graph, the CMC would again be trivially satisfied and provide no additional constraints or information about the joint probability distribution of the variables in the graph.[9]

Thus, some additional principle is needed to form substantive inferences between statistical data and causal graphs. The other condition that has most frequently been adopted to play this complementary role is the Causal Faithfulness Condition (CFC).[10] The CFC has received less philosophical attention, though it is generally regarded as the less secure of the two conditions by

[9] It should be noted, however, that the CMC is typically only applied in the context of directed acyclic graphs (DAGs). And since the situation where every variable is a direct cause of every other variable in a graph would entail the presence of causal cycles, this situation would be ruled out by definition in the context of DAGs.
[10] See also the notion of 'stability' in Pearl (2000).

proponents and developers of causal discovery algorithms (Zhang and Spirtes 2008). The CFC is the converse of the CMC. Informally, it states that causation implies correlation (probabilistic dependence)—i.e. all variables that are causally related should be probabilistically dependent. More formally, the CFC states that the independences in a probability distribution for a set of variables are only those entailed by the CMC. So, if two variables are probabilistically independent, there is no causal relation between them. Thus, where the CMC stipulates a sufficient condition on two variables being causally linked, the CFC stipulates a necessary condition—i.e. it is necessary that they share a correlation that cannot be screened off.

The CFC is perhaps less intuitive as a principle relating causation and probability. On one hand, it is intuitive that a causal path between two variables should imply a dependence between those variables. But what this claim is effectively ruling out (as will be discussed in more detail below), is the possibility of canceling causal pathways between two variables. There is nothing inherent in our conception of causation and how causation implies correlation that rules out canceling causal pathways. The CFC is, however, a very powerful complement to the CMC. Jointly, for any causal graph, the CMC and CFC entail a complete set of probabilistic relationships, both conditional and unconditional dependencies and independencies, between all variables in that graph (though that set of probabilistic dependences may be compatible with more than one causal graph). With the addition of the CFC, the concerns about trivial satisfaction of the CMC mentioned above are eliminated.

Like the CMC, the CFC contains information relevant to both causal hypothesis testing and causal discovery. With respect to causal hypothesis testing, the CFC entails that all variables in a causal graph connected by a causal pathway are probabilistically dependent. With respect to causal discovery, the CFC entails that any independence between two variables, conditional on all other

variables in the graph, implies that there is no causal pathway between those variables. In other words, the CFC licenses inference from probabilistic independence to the absence of a causal relationship between variables. (Again, see *Table 1* for reference.)

Table 1: Breakdown of inferences supported by the CMC and CFC.

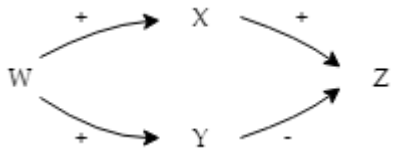
	Causal discovery	Causal hypothesis testing
CMC	If $\sim(C_i \perp\!\!\!\perp E_j)$, then C_i causes E_j	If C_i does not cause E_j, then $C_i \perp\!\!\!\perp E_j$
CFC	If $C_i \perp\!\!\!\perp E_j$, then C_i does not cause E_j	If C_i causes E_j, then $\sim(C_i \perp\!\!\!\perp E_j)$

Note: Here $\perp\!\!\!\perp$ means "is independent of, conditional on all other variables in the causal graph."

However, there are multiple ways the CFC can be violated. For instance, in deterministic systems, if X is a deterministic cause of Y and the true causal DAG were X→Y→Z, Y and Z would be independent conditional on X. Further, if a variable in a DAG were to maintain the same parameter value for a particular data set, it would be rendered probabilistically independent of its descendants in the joint distribution function for that data set. For instance, suppose (quite plausibly) that gender is a cause of propensity to commit gun violence. These two variables would be probabilistically independent in a sample that only included men. Neither of these forms of violation have been the focus of debates about the CFC. With respect to the former, this is because the types of systems these causal modeling techniques are applied to tend to be nondeterministic, and fully deterministic causal relationships would be discoverable by other means. With respect to the latter, such a bias in a variable of interest in a data set would typically be very apparent to researchers interested in causal relationships in such a system. Moreover, this scenario is ruled out by the assumption of positive support—i.e. that all values of the variables have positive probability—which is commonly made in the context of GCMs.

There is, however, a separate set of cases in which the CFC is violated that are potentially more problematic for the typical domain of application of GCMs. Specifically, violations of the CFC occur when there are precisely balanced inhibitory (suppressive) and excitatory (stimulative) causal pathways to some effect variable. For example, consider a triangular causal structure with a direct path A→C and another path A→B→C. Suppose the strength of the excitatory pathway from A→C were exactly equal to the net inhibitory effect of the pathway A→B→C. The result would be that the causal relation between A and C would be masked in the probability distribution for this system. That is, A and C would be probabilistically independent despite the presence of a causal relationship between them—thus exhibiting a violation of the CFC.

For a more concrete example, consider the following. Consumption of fish increases both omega-3 fatty acid levels and levels of mercury in the blood (Chowdhury et al. 2012; Cole et al. 2004). Mercury increases risk of dementia (Hock et al. 1998); omega-3 fatty acids decrease risk of dementia (Lim et al. 2005). If, in a particular population, the negative effects of levels of mercury absorbed through fish consumption precisely balanced the positive effects of omega-3 fatty acids due to fish consumption (relative to their effects on dementia), then there would be no net effect of fish consumption on the risk of developing dementia. Thus, development of dementia would be probabilistically independent of fish consumption even though fish consumption is causally relevant to development of dementia. Such a case would constitute a violation of causal faithfulness.



Linear causal model (excluding error terms): X=aW, Y=bW, Z=cX+dY

Figure 4: Example of a possible structure exhibiting a violation of faithfulness.

Note: From the example above, consider W=levels of fish consumption, X=levels of mercury, Y=levels of omega-3s, and Z=risk of dementia. Note that a failure of faithfulness requires precise balancing between the excitatory and inhibitory pathways—given the proposed linear causal model with the figure, this will occur only when ac+bd=0.

The controversies that have surrounded the CFC have generally focused on how likely such precisely balanced pathways are to occur and hence how likely the CFC is to fail in this manner.[11] In the example in *Figure 4*, it is clear that such a "conspiracy of the evidence" would be highly unlikely. In fact, and to be clear, the benefits of omega-3 fatty acids associated with fish consumption seem to far outweigh any risks due to increased mercury levels associated with fish consumption (Mutter et al. 2007). This unlikeliness of precisely balanced pathways has been the primary route to defending or justifying the use of the CFC in causal inference. In particular, Spirtes, Glymour, and Scheines (2000), henceforth SGS, provide a proof that, for linear causal models, the set of parameter values that correspond to violations of faithfulness is Lebesgue measure zero compared to the set of all possible parameter values.[12]

To cash this out, grant that a causal graph with n parameters will have a set of possible parameter values corresponding to an n-dimensional real space ($\mathbb{R}^n$), which is to say each parameter can take any real number value. Any subset of that space that corresponds to an (n-1)-dimensional space is said to be Lebesgue measure zero with respect to $\mathbb{R}^n$. To perhaps make this more intuitive, consider a cube, out of which you slice a single square. If you wanted to know how much of the volume of the cube is contained in that square, the answer would be zero. This is, by

[11] More on this to follow in §II.
[12] See Theorem 3.2, SGS pp.41-42; and pp.383-384 for the proof; Meek (1995) extends this theorem to causal models with discrete variables.

extension to any number of dimensions, what it means to say that a set is Lebesgue measure zero with respect to a superset. What SGS show is that the subset of possible parameter values that constitute a violation of faithfulness has at least one less dimension of variation than the set of all possible parameter values. We can again see this more intuitively with the example from *Figure 4*. Allow that the set of possible parameterizations of this causal model correspond to a 4-dimensional real space. The subset of that space corresponding to violations of faithfulness is those parameterizations for which ac+bd=0, which corresponds to a 3-dimensional space (note that once any three parameters are set, the fourth is determined).

From this proof, SGS (and others) claim that because violations of faithfulness are measure zero, the probability of violations of faithfulness should also be zero, thus justifying the use of the CFC as a principle for inference between causal graphs and quantitative relationships between variables in those graphs. SGS's argument can be reconstructed as follows:

> P1: The set of all possible combinations of parameters values for a causal model with N parameters constitutes an n-dimensional real space ($\mathbb{R}^n$).
>
> P2: The subset of that space that corresponds to violations of faithfulness is (n-1)-dimensional or less.
>
> P3: Any subset of $\mathbb{R}^n$ that is N-1 dimensional or less is Lebesgue measure zero.
>
> P4: Any subset of $\mathbb{R}^n$ that is Lebesgue measure zero has probability zero of occurring.
>
> C: Therefore, violations of faithfulness have probability zero of occurring.

P1 and P3 are uncontroversial. Establishing P2 is the main focus of SGS's analysis, and is well supported, as can be seen from the discussion of the example from *Figure 4*. P4 is left implicit in SGS's original argument and deserves further scrutiny.

3.3 The Likelihood of Failures of Faithfulness

The line of justification for the CFC offered by SGS has not gone unchallenged. Several authors have argued, despite SGS's proof, that the CFC is likely to fail in a range of cases that all fall within the typical domain of application of DAGs. For instance, Nancy Cartwright (1999a, 1999b, 2007) has argued that violations of the CFC will be common in certain engineering, economic, and medical contexts. She argues,

> It is not uncommon for advocates of DAG-techniques to argue that cases of cancellation will be extremely rare, rare enough to count as non-existent. That seems to me unlikely, both in the engineered devices that are sometimes used to illustrate the techniques and in the socioeconomic and medical cases to which we hope to apply the techniques. For these are cases where means are adjusted to ends and where unwanted side effects tend to be eliminated wherever possible, either by following an explicit plan or by less systematic fiddling. The bad effects of a feature we want - or are stuck with - are offset by enhancing or encouraging its good effects. (Cartwright 1999, p. 16)

> A good many of the systems to which we think of applying the methods advocated by Bayes-net theorists are constructed systems, either highly designed… or a mix of intentional design, historical influence and unintended consequences, as in various socio-economic examples. In these cases cancellations of the effects of a given cause, either by encouraging the action of other factors or by encouraging the contrary operation of the cause itself, can be an important aim, particularly where the effect is deleterious. (Cartwright 2007, pp.70-71)

Kevin Hoover has made a similar argument for the likelihood of violations of faithfulness in macroeconomic contexts where policymakers aim to precisely balance parameters to ensure stable outcomes. He writes,

> [SGS] acknowledge the possibility that particular parameter values might result in violations of faithfulness, but they dismiss their importance as having 'measure zero'. But this will not do for macroeconomics. It fails to account for the fact that in macroeconomic and other control contexts, the policymaker aims to set parameter values in just such a way as to make this supposedly measure-zero

> situation occur. To the degree that policy is successful, such situations are common, not infinitely rare. (Hoover 2001, p.170)

Finally, Holly Andersen has applied a similar line of reasoning to biological systems that maintain equilibria. She writes, "some kinds of systems, especially those studied in the so-called special sciences, are likely to display the kinds of features that lead to CF violations, such as mechanisms for equilibrium maintenance across a range of variables" (Andersen 2013, p.682).

All these arguments effectively challenge the claim that measure zero subsets of $\mathbb{R}^n$ will have zero probability of occurring—i.e. P4 in the reconstruction at the end of §I. Both Cartwright (2007) and Andersen (2013) are explicit about this:

> But this conclusion would follow only if there were some plausible way to connect a Lebesgue measure… with the way in which parameters are chosen or arise naturally for the causal systems that we will be studying… [W]e not only need a story that connects a Lebesgue measure… with how real parameter values arise, but we need a method that selects as a question to be addressed before values are chosen: shall values occur that satisfy faithfulness or not. (Cartwright 2007, p.68)

> However, the fact that CF-violating systems are measure 0 in this class does not imply that we will not encounter them with any frequency… [Rational numbers] are also measure 0 with respect to the real numbers, while irrational numbers are measure 1… However, this does not imply that rational numbers are unlikely to be encountered simpliciter: bluntly put, we do not encounter numbers by randomly drawing them from the number line. Rational numbers are encountered, and used, overwhelmingly more often than one would expect from considering only the proof that they are measure 0 with respect to the real numbers. (Andersen 2013, p.677)

The crux of these likelihood objections rests on this point—the connection between Lebesgue measures and the probabilities of particular sets of parameter values arising. In their original presentation of the measure zero argument, SGS offer no explicit justification for P4—instead merely offering the 'measure zero' proof and stating that this implies that failures of the CFC are

probability zero.[13] Other authors have, however, attempted to further develop a rationale for P4 that Cartwright, Hoover, and Andersen's arguments do not take into account.

Woodward (1998) and Pearl (1998) in separate commentaries on SGS's work, home in on the assumptions about the parameters in a causal model that would have to hold for P4, and thus the CFC, to be justified.[14] Woodward encourages caution in the conditions under which faithfulness is employed. In particular, he notes that other theoretical information is often necessary to provide context for evaluating causal claims, pointing out that there are familiar and well-established cases where parameter values precisely cancel in ways that "mask" causal relationships (Woodward 1998, pp.142-145). For instance, consider a particle at rest in Earth's gravitational field. One explanation for why the object is at rest is that there are no forces acting on it; another explanation is that there are a number of forces acting on it in a way that is precisely balanced such that the object remains at rest. The CFC rules out the latter in favor of the former, but obviously that is not the better explanation in this case, because we have independent theoretical knowledge that all objects in Earth's gravitational field are subject to a constant gravitational force that must be counterbalanced for an object to be at rest. Woodward concludes from this example that "[e]xplanations that eschew special parameter values and complicated causal influences are not always preferable to those that do not" (Woodward 1998, p.144). This provides a useful reminder that the CFC should not be taken as an exceptionless condition on causal claims and causal

[13] In later work, Spirtes et al. (2004) offer some further elaboration, "[s]ome form of assumption of faithfulness is used in every science, and amounts to no more than the belief that an improbable and unstable cancellation of parameters does not hide real causal influences" (Spirtes et al. 2004, p.182). However, here again the improbability of cancellation of parameters is merely stated. The notion that violations of faithfulness are unstable would seem to come from the idea that, because the pathways must be so precisely balanced, any slight deviation in one or more parameter would reveal the 'hidden' causal relationship. Woodward (1998) expounds on this notion of instability.

[14] Specifically, these commentaries were directed at the TETRAD Project (Scheines et al. 1998), though the themes are not specific to that piece and have implications for the broader SGS program.

explanations, but instead as a heuristic to be applied only in particular circumstances in which it is likely to hold.

The question then becomes what factors bear on the appropriateness of applications of the CFC. To this end, Woodward points out that the line of reasoning in the gravity example may not apply to the social sciences and other fields more typically within the domain of DAG modeling techniques. For instance, Glymour et al. (1987) offer an analysis of a large-scale social experiment that tested the influence of monetary disbursements, administered through the Transitional Aid Research Program (TARP), on recidivism, as discussed by Rossi et al. (1980). The experiment found no effect between recidivism and monetary disbursements—i.e. the treatment group that received disbursements had recidivism rates that were not significantly different from recidivism rates in the control group that did not receive disbursements. A natural interpretation, as is implied by the CFC in this case, is that there is no causal influence between monetary disbursements and recidivism. Rossi et al., however, postulated that TARP payments decreased recidivism rates, but that they also increased unemployment, which led to a corresponding increase in recidivism that precisely counterbalanced the decrease directly resulting from TARP payments. Following Hans Zeisel (1982), who was on the advisory board for the TARP study and disagreed strongly with this interpretation, Glymour et al. argue that this is a case where the CFC holds, and so Rossi et al.'s model, with its unnecessarily complicated causal structure that requires special parameter values, should be rejected.[15]

There are two notable differences between this case and the gravity example. First, there is vast divide between the independent theoretical motivations at play in the gravity example as compared to Rossi et al.'s proposed causal structure and the special parameter values it requires.

[15] See Zeisel 1982 and Glymour et al. 1987, pp. 26-30.

In the gravity example, the independent theoretical motivation is the primary factor directing us toward the correct conclusion that this is a case in which special parameter values, supporting precisely balanced pathways, are warranted. Rossi et al.'s proposed model was not completely unmotivated. They related their results back to a smaller previous study, the Baltimore LIFE experiment, which had found an 8% reduction in arrests for crimes of theft (but no effect for other crimes) for those receiving monetary disbursements compared to controls (Rossi et al. 1980, pp. 37-43). It is a stretch to use a result from a smaller study that is barely statistically significantly and is restricted to a single type of rearrest charge as the linchpin for the central interpretation of results in the TARP study.[16] At any rate, it suffices for present purposes to note that there is a wide gulf between the theoretical motivation at play in the TARP and that motivating the causal story in the gravity example.

The second and related difference is that the relevant parameters in the TARP study do not operate as physical constants, but instead are relatively "unstable", making the special parameter values and precisely balanced pathways seem less plausible. Woodward further notes that often in the social sciences (qualitative) causal structure may be stable despite (quantitative) variation in the parameter values across different instances. For instance, the economic principle of supply and demand seems to express a stable qualitative causal relationship; however, the particular quantitative relationships that exhibit supply and demand can vary dramatically. In such contexts, the parameters in causal models will be not behave like the physical constants at play in the gravity example. As a result, Woodward argues, it becomes more reasonable to assume that precisely balanced causal pathways will not occur. Thus, in conditions where the underlying qualitative causal structure is stable, and the parameters are relevantly unstable, the CFC will be more useful

[16] Again, see Zeisel 1982 and Glymour et al. 1987, pp. 26-30, for further discussion of this case.

as a constraint on causal modeling. Summarizing this rationale, he writes, "[a]ssuming that the structural coefficients are unstable and free to change independently of each other, one might argue that models which imply conditional independence relations because of special coefficient values should be rare and unlikely to persist across time and space" (Woodward 1998, p.145). In other words, violations of faithfulness should be rare and unstable when parameters are unstable and vary independently of one another.

Pearl (1998) offers a similar rationale, stressing the importance of the notion of "autonomy" (Aldrich 1989), also known as modularity (see chapter 4 for additional discussion), in justifying faithfulness. Pearl explains, "[t]his invariance means that mechanisms can vary independently of one another, which implies that it is the set of structural coefficients… rather than other types of parameters, that will vary independently when experimental conditions change" (Pearl 1998, p.121). Both Woodward and Pearl thus see the ability of parameters to vary independently as crucial to justifying the CFC. Woodward adds that parameters need to be unstable and also that there need to be no independent theoretical reasons to expect precisely balancing forces (as in the gravity example). Given that these conditions are met, it is reasonable to expect that P4 will hold, and thus that violations of the CFC will be probability zero.

Steel (2006) goes further, offering a more explicit and precise analysis of the conditions under which P4 holds. To understand Steel's analysis, we first need a general way of assigning probabilities to subsets of $\mathbb{R}^n$. To this end, we can associate each parameter of a causal model with a random variable. The set of all random variables for a causal model with n parameters is then V = {V1, V2, … Vn}. The question becomes, what conditions have to hold of the joint probability distribution of V for P4 to hold? That is, what would the joint distribution function need to look like for it to be the case that subsets of $\mathbb{R}^n$ of Lebesgue measure zero receive zero probability.

With this setup, we can recast Woodward and Pearl's claims that the CFC will be appropriate when parameters are allowed to vary independently. For a parameterization to be variation independent is for the parameter space to be the Cartesian product of the sets of values of each of the parameters.[17] In other words, no particular parameter or group of parameters constrains the value that any other parameter can take, so all possible combinations of parameter values are in the parameter space. Steel argues that variation independence by itself is neither necessary nor sufficient to rule out strict exceptions to the CFC (by justifying P4). On one hand, if one or more of the random variables in V is discrete, Lebesgue measure zero subsets of $\mathbb{R}^n$ will have non-zero probability regardless of whether the parameters are variation independent.[18] A variable's being discrete just means that it has positive probability for at least one point value (whereas continuous variables have positive probability only for some range of values and have zero probability for all point values).

Take the case where n=1 as an illustration. Any point value is measure zero with respect to the real number line. So, if our single random variable is discrete, a measure zero subset of $\mathbb{R}$ receives positive probability. As will become clear in what follows, this generalizes to any number of variables—one or more variables being discrete collapses the number of dimensions of $\mathbb{R}^n$ that receive positive probability regardless of whether those variables are variation independent. Thus, variation independence by itself is insufficient. An additional condition is required—namely that the marginal distributions of all random variables associated with the parameters be continuous. In other words, every individual parameter must be associated with a continuous random variable,

[17] See, e.g., Lindsey (1996) and Bergsma and Rudas (2002).
[18] See Steel (2006), p.310 for a counterexample illustrating this point. But as will be discussed in the remainder of this section, any variable being discrete would obviously entail a failure of joint continuity, and joint continuity is necessary for P4.

which just means that the parameter can take any real value within some interval. For ease of terminology, I will refer to this requirement simply as the requirement that the parameters be "marginal continuous." If both these conditions are met—if the joint probability distribution of V satisfies both marginal continuity (of all variables) and variation independence—then subsets of $\mathbb{R}^n$ of Lebesgue measure zero will have zero probability.

This deserves some additional unpacking. The probability associated with a continuous random variable corresponds to the area under the curve of the variable's probability density function, where the total area under the curve is 1. Critically, only ranges of values receive positive probability for continuous random variables—point values receive zero probability (as there is no area above a point). For a single variable, the probability density function thus specifies an area over a line, which corresponds to the relevant range of values of the variable. A variable fails to be continuous (i.e. is discrete) insofar as it has positive probability at a particular point.

Joint continuity extends this concept to a set of variables. For two random variables the probability density function specifies a volume over a plane, which corresponds to an area equal to the product of the two relevant ranges for each variable. Again critically, only ranges of values for both variables receive positive probability (as there is no volume above a line). It becomes more difficult to visualize with three or more variables, but the same principles iterate for any number of variables. Thus, for a set of random variables to be jointly continuous, there must be some joint probability density function that specifies probabilities for a set of ranges of each variable. So, for a set of n random variables, the joint probability density will be an (n+1)-dimensional "curve" over an n-dimensional space, which corresponds to the product of the relevant ranges of each variable. Only subsets of $\mathbb{R}^n$ with n-dimensions receive positive probability. Any

subset of $\mathbb{R}^n$ that has (n-1)-dimensions or fewer receives zero probability, which is to say any Lebesgue measure zero subset of $\mathbb{R}^n$ receives zero probability.

The existence of a joint probability density function for a set of variables requires not only that the marginal distribution of each variable be continuous, but also that probability density function of each variable be continuous conditional on any subset of the others. These conditions are satisfied by the conjunction of marginal continuity and variation independence. If the marginal distributions of all variables are continuous and no particular variable or group of variables constrains the value of any other variable, then the probability density functions of all variables will be continuous conditional on any subset of the others. Thus, marginal continuity in conjunction with variation independence entails joint continuity which in turn entails P4—that any Lebesgue measure zero subset of $\mathbb{R}^n$ receives zero probability. This delivers a more precise version of Pearl and Woodward's claim that the CFC may hold when parameters are unstable and vary independently. This claim is true insofar as the required instability entails marginal continuity, as marginal continuity and variation independence are jointly sufficient for P4.

However, as Steel notes, variation independence, in the technical sense, is stronger than necessary: "joint continuity does not require that variation independence be true, since the range of possible values of one variable may be restricted by the value of another even if each variable is continuously distributed conditional on any combination of other variables" (Steel 2006, p. 311). Thus, we can replace the requirement that no particular parameter or group of parameters constrains the value that any other parameter can take with the requirement that all parameters vary continuously conditional on any other parameter or set of parameters. Steel ultimately argues that this condition is reasonable in the typical domain of application for DAG causal modeling because "it is quite plausible that this is indeed the case in biology and social science, and indeed,

in any field that studies complex systems in which the strength of causal relationships depend on a plethora of variable factors" (Steel 2006, p. 311). However, as I'll argue in the next two sections, this may fail to be true in an important set of cases in neuroscience that generalize in relevant ways to other complex systems.

3.4 Failures of Faithfulness in Neuroscience

Robustness in neural systems provides a useful testing ground for these points of debate regarding the likelihood of violations of the CFC. I argued in chapter one that neural robustness provides clear cases of causal explanatory multiple realization (CEMR)—i.e. relevant similarity in function despite relevant variation in the causal mechanisms that give rise to that function. The connection between CEMR and violations of the CFC is straightforward to demonstrate. Let both the relevant aspects of the mechanism(s) supporting a function and the function itself be nodes in a causal graph. In the true causal graph of such a system, there will be a number of directed edges connecting the relevant aspects of the mechanism to the function in question. In cases of robustness, there will be stability in function—circumstances in which the value of the function variable remains fixed—despite variation in the values taken by causal variables. To the extent that this occurs, there will be probabilistic independence between the function variable and some set of its causes. This is a violation of the CFC—probabilistic independence despite the presence of causal connection.

This can be illustrated more precisely through examples of robustness of neural function at the single cell level. The functions of individual neurons are generally characterized in terms of their response (input-output) properties. A neuron's response properties are determined by the

combined effect of several currents that result from proteins (ion channels) allowing ions to permeate the cell's membrane. Whereas a neuron may live for decades, these proteins have shorter lifespans, typically turning over on the order of hours, days, or weeks.[19] As a result, the conductances that determine a neuron's response properties also vary over time. This raises questions regarding how stable those response properties are, and, to the extent that they are stable, how neurons maintain that stability. A range of studies has shown, for a variety of cells, that ion channel densities can in fact vary severalfold between cells that nonetheless have effectively identical response properties (Golowasch et al. 2002, Schulz et al. 2006, Ransdell et al. 2013). This suggests that neuronal response properties are tightly regulated and has motivated research into the mechanisms of that regulation.

Burst firing in Purkinje cells is a prime example of this sort of robustness of response properties to variation in the underlying conductances. Purkinje cells are a type of neuron found in the cerebellum that are relatively large cells with sprawling dendritic trees that receive tens of thousands of inputs. They play key roles in motor behaviors and, particularly, in motor learning. Climbing fibers, projections from the inferior olivary nucleus in the medulla oblongata, provide a strong source of excitatory input to Purkinje cells. According to a longstanding model, the inputs from climbing fibers convey a motor error signal that is integral to motor control and motor learning.[20] Depolarizing stimulation from climbing fibers evokes a stereotyped all-or none burst firing pattern from Purkinje cells. This burst firing is a crucial function that enables plasticity in adjacent circuits, in the form of both long-term potentiation and long-term depression.

---

[19] See, e.g. Hanwell et al. 2002; and Marder and Goaillard 2006, for relevant review.

[20] This is the Marr-Albus-Ito model, one of the most influential computational cum experimental models in neuroscience (Marr 1969, Albus 1971, Ito et al. 1982, Ito and Kano 1982, Ito 1989). For relevant review of the model, see Strata 2009. For a recent extension of this model that proposes that this cerebellar circuitry constitutes a more general reinforcement learning mechanism, see Yamazaki and Lennon 2019.

Purkinje cell burst firing is the result of a relatively small net inward current after an initial action potential. And Purkinje cells are among those that have been shown to have similar electrophysiological profiles despite massive variance in surface conductances. That is to say, Purkinje cells with very similar burst profiles exhibit substantial variation in underlying conductances of different types of ions. This similarity in burst profiles is surprising because the bursts are triggered by small net influx of currents relative to the variability in any particular conductance.

Imagine a water lock designed such that there are multiple gates instead of one between any two chambers, where each gate has a pump that pushes water into one chamber or the other. Now imagine that when these gates open and the pumps turn on, huge amounts of water flow between the two chambers. However, the result when the gates close is the same—a small amount of water, relative to the large fluxes, always enters the downstream chamber. Such a system would be very sensitive to the size of the gates, strength of the pumps, etc. Varying the size of any single gate would throw off the balance and change the net amount of water flowing between the two chambers. This is analogous to how burst firing is triggered in Purkinje cells—multiple large inward and outward currents sum to a relatively small net inward current after an initial action potential, in turn triggering further action potentials. It would seem that small changes in those large inward and outward currents would be likely to throw off the balance and disrupt bursting and yet that is not what is observed (Swensen and Bean 2003).

In a remarkable study, Swensen and Bean (2005) investigated the mechanisms that support robustness in Purkinje cells. They performed two distinct interventions that targeted different timescales: (1) pharmacological blockade of sodium conductance, (2) genetic knockout of sodium ion channels. Pharmacological blockade is transient and occurs on very short timescales. The short

duration of the intervention rules out any second messenger processes occurring within the cell to alter ion channel expression on the cell membrane. Genetic knockout persists through the life of the organism, thus allowing second messenger processes to potentially alter ion channel expression. Under both interventions, burst firing was surprisingly robust, and Swensen and Bean's additional analysis revealed that there are in fact distinct mechanisms that support robustness on the two different timescales. For present purposes, I will be concerned only with the pharmacological knockout aspect of Swensen and Bean's study, though the genetic knockout experiment, as well as the significance of the distinction between the two forms of intervention, will be a focus of discussion in chapter four.

In their pharmacological intervention, Swensen and Bean used a substance called tetrodotoxin (or TTX) to temporarily block sodium conductance via voltage-gated sodium channels. The vast majority (70%) of cells persisted bursting with a 50% reduction in sodium conductance.[21] In a previous study, Swensen and Bean (2003) had found that TTX-sensitive sodium current contributed the largest of the inward currents during the interspike interval. Swensen and Bean's (2005) result thus demonstrates a remarkable stubbornness of Purkinje cells to persist in bursting. Returning to the water lock metaphor, it would be as though the same net amount of water made it into the downstream chamber despite varying the size of the largest gate by 50%. (For perspicuity of the metaphor, recall that the largest gate allows an amount of water through that dwarfs the net amount of water that passes through after the gates have closed.) Within that range of variance, the gate size would be rendered independent of the amount of water that reaches the downstream chamber. This would, in ways that will be explored further in what follows, pose serious challenges

[21] With a 25% reduction in sodium conductance, all the Purkinje cells persisted bursting; with a 50% reduction, the majority (70%) of cells continued bursting; it was only with a 75% reduction that all cells ceased bursting.

to understanding the causal dynamics of the system. Similarly, for Purkinje cells, within the 50% range of TTX blockade, changes in sodium conductance are rendered independent of Purkinje cell burst firing, despite being a relevant cause.[22] In both cases, prior knowledge of the causal structure of the system, specifically the knowledge that these factors—gate size, or sodium conductance—*must* be causally relevant factors, invites further investigation into the means by which the robustness is achieved. That is, because the causal structures of these systems are already somewhat characterized, the robustness over variation in causally relevant factors suggests the presence of precisely balanced pathways.

Swensen and Bean investigated the means by which Purkinje cells achieve robustness to TTX blockade. The main currents that determine burst firing are sodium, calcium, and potassium currents. Each of these ions can have a number of associated ion channels that regulate the flow of ions across the cell membrane. Because TTX has transient effects and operates acutely, and bursting was assayed within seconds, it is essentially impossible that any preservation of burst firing would be due to second messenger systems that affect the expression of channels that regulate these currents. That is, changes in ion channel density can be ruled out as the source of robustness on theoretical grounds. Particular ion channel types, however, do have their own

[22] Note that this example involves independence for a range of values (0-50% TTX blockade), which is typical for examples of robustness. Within the GCM framework, this kind of independence is consistent with there still being a discoverable causal relationship between the variables at issue (in this case burst firing and TTX-sensitive sodium conductance). In other words, in GCMs the only thing that is required for X to cause Y is that Y is dependent on some values of X in some background conditions. Thus, in examples of robustness, perturbing the system outside the relevant range (in this case >50% TTX blockade) will reveal dependence between the variables. The issues raised by robustness might thus be better characterized as "local failures" of the CFC or as invariance over a certain set of values that a variable can take. However, such failures are no less relevant for several reasons. Most notably, we are often interested in discovering causal structure within such ranges. In the particular case of Purkinje cells, the 0-50% range corresponds to a generous approximation for "normal operating conditions" of a cell—perturbation beyond that range is not something that would occur in vivo. So, if we are interested in discovering causal structure within such ranges, invariance to all values a variable can take in that range will pose problems for causal discovery that mirror the problems associated with general failures of the CFC.

dynamics, and changes in the parameters governing those dynamics are the obvious additional place to look for compensatory changes.

To determine which other ion channels might be involved in the preservation of bursting, Swensen and Bean first recorded the action potential waveforms for each Purkinje cell in current-clamp, and then switched to voltage-clamp and played the action potential waveform back to the cell.[23] This allowed them to observe any compensatory changes in other conductances that facilitated continued burst firing. They discovered that the acute decrease in sodium conductance due to TTX produced a decrease in the height of the action potential and a hyperpolarizing shift in postspike membrane potential. These changes in action potential waveform effected a small decrease in calcium conductance and, more notably, significant reductions in potassium current—both voltage-dependent and calcium-activated potassium current saw reductions ~50% in comparison to their levels with no TTX block. For simplicity, we can neglect the relatively small reduction in calcium conductance and represent the causal structure of the basic compensatory mechanism Swensen and Bean discovered as follows in *Figure 5.*

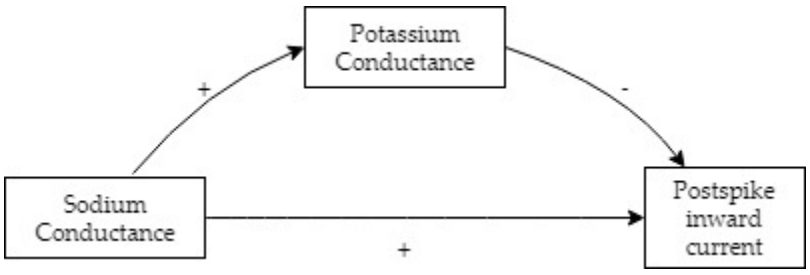


Figure 5: Simplified causal DAG for Purkinje cell burst firing robustness.

[23] Current-clamp is a technique in which experimenters record from cells while injecting stimuli (usually 1ms depolarizing electrical currents) and recording the cell's response (so the cell's membrane potential is the dependent variable). Voltage-clamp, by contrast, is a technique in which the experimenter controls cell's membrane potential and observes how aspects of the cell's electrophysiology respond (so the cell's membrane potential is the independent variable).

Note: As sodium conductance decreases, potassium conductance (through two channels—voltage-dependent and calcium-activated) also decreases. The net effect balances to preserve the small net postspike inward current that drives bursting.

Sodium exists in high concentrations outside the cell, whereas potassium exists in high concentrations inside the cell. Thus, when TTX-sensitive sodium channels open, sodium rushes into the cell, depolarizing the membrane; and when voltage-activated and calcium-activated potassium channels open, potassium rushes out of the cell, hyperpolarizing the membrane.

Both voltage-dependent and calcium-activated potassium channels are sensitive to changes in the action-potential waveform. The primary driver of decrease in voltage-dependent potassium is the hyperpolarizing shift in interspike potential. Swensen and Bean hypothesized that this decrease is due to the voltage sensitivity of deactivation of the Kv3-type potassium channel in the observed voltage range. They attribute the decrease in calcium-activated potassium current to three main factors: (1) the decrease in calcium entering the cell (despite the changes in calcium current being relatively small), and the hyperpolarizing postspike shift serves to (2) promote deactivation of BK-type potassium channels, and (3) decrease the potassium driving force. The causal dynamics here are obviously fairly complex, which on one hand makes it all the more remarkable that these pathways wind up effectively canceling. On the other hand, the complex causal dynamics combine to produce a fairly simple causal structure (*Figure 5*). Intervening on sodium conductance disrupts the main inward current during the interspike interval. However, it also affects the action potential waveform (it is not possible to intervene on TTX-sensitive sodium conductance without also changing the action potential waveform). This leads to a form of *parameter coupling* whereby potassium conductance, which is the main source of outward current flow in the interspike interval

(Swensen and Bean 2003), also decreases as a result of the TTX treatment.[24] The net effect is that the postspike inward current is stable and burst firing persists.

This basic causal structure mirrors a classical violation of faithfulness.[25] And the parameter coupling governing the effects of sodium conductance and potassium conductance during the interspike interval reveals this as, in fact, a fairly clear case of a failure of the CFC. In this case the effect variable, net postspike inward current, is probabilistically independent of variation in sodium conductance, even though sodium conductance is clearly a cause of that postspike current. This example thus provides a clear case of a violation of causal faithfulness with a reasonably well characterized causal mechanism underlying that violation.

Before moving on to consider how such failures of the CFC can occur (given the arguments considered in §II), it is worth stressing that the specific example of robustness in Purkinje cells is hardly unique in neural systems. As mentioned at the beginning of this section, many other neuronal types have been found to exhibit severalfold variation in intrinsic conductances, similar to the variances found in Purkinje cells (Golowasch et al. 2002, Schulz et al. 2006, Ransdell et al. 2013). This only makes sense, as all neurons face the same basic problem of maintaining stable functions despite the components supporting those functions changing over time. As Marder and Goaillard (2006) state in their review of work on this topic,

> [E]ach neuron is constantly rebuilding itself from its constituent proteins, using all of the molecular and biochemical machinery of the cell. This allows for plastic changes in development and learning, but also poses the problem of how stable

[24] The sort of parameter coupling at issue here indicates a stable functional relationship between these parameters in the system. In their original formulation of the measure zero argument, it is clear that SGS thought of different combinations of parameter values as "independent draws" and thus that any cancellations would be accidental and not the result of stable functional relationships.

[25] This basic causal structure provides, in fact, a so-called "triangle failure of faithfulness" (Zhang and Spirtes 2008). The significance of triangle failures of faithfulness is that they are not "detectable" from data alone, where non-triangle failures of faithfulness are. That is, non-triangle failures of faithfulness have intervening variables that could, in principle, be revealed through appropriate experimental interventions.

> neuronal function is maintained as individual neurons are continuously replacing the proteins that give them their characteristic electrophysiological signatures. (Marder and Goaillard 2006: 563)

In other words, the problem that generates the need for robustness in Purkinje cells is a general problem for all neurons (and indeed is faced by all complex systems where stability of function must be achieved despite dynamic components). In Purkinje cells and other neurons, this stability of function seems to be achieved via compensatory differences in other causally relevant features—either through variance in other intrinsic conductances or through compensatory changes in ion channel density.

And indeed, computational models for other types of neurons demonstrate that a variety of intrinsic conductances and ion channel densities can give rise to similar electrophysiological profiles (Goldman et al. 2001, Taylor et al. 2009, Ball et al. 2010). This suggests that compensatory mechanisms, like those discovered by Swensen and Bean, that maintain target levels of activity, will be found throughout the nervous systems of humans and other animals.[26]

Similar considerations scale up to the level of small networks in neural systems. Such networks can be perturbed by either removing individual neurons or by varying the properties of those neurons. For long-lived organism, the problem more commonly is the latter—i.e. variability in the conductances of the neurons that comprise the network. The reason, as above, is that neurons tend to persist over long periods of time—from decades to the lifespan of the organism. Since the relevant variance is thus at the level of components of the neurons that compose the network (in essence, two "levels" down from the function in question), one might suppose that the natural way to ensure stable output of the network is just to tightly regulate the output of the individual cells

[26] Biological evidence of such compensatory mechanisms has indeed been found for a number of different neuronal types—see, e.g., MacLean et al. 2003, Guo et al. 2005, Nerbonne et al. 2008.

via the same compensatory mechanisms mentioned above. However, another way to solve this problem is to reconfigure the network, adjusting the response properties of other cells within the network to compensate for changes in the response properties of some particular neuron. That is, instead of tightly regulating the roles that each individual cell plays within the network, compensation can occur by adjusting the roles played by other cells in the network. The pyloric network of the stomatogastric ganglion, discussed in chapter one, is an example of this later sort of compensation (Prinz et al. 2004). Similar studies have demonstrated the same principle in other small networks (Goaillard et al. 2009, Grashow et al. 2010, Ransdell et al. 2012).

Scaling up to larger networks, the situation becomes more complex. This is primarily due to vast number of potentially relevant parameters. Computational studies, like those showing the multitude of combinations of parameter values that can support stable function in single neurons and small networks, are simply not feasible for larger networks. There are, however, compelling reasons to again believe that similar principles will hold—i.e. that network function will exhibit stability over large variations in the parameters that influence network activity.

In the first place, the same problem exists insofar as large-scale networks are comprised of small networks and neural circuits, which are in turn comprised of individual neurons. Those individual neurons are in the same state of continuous flux that induces robustness on smaller scales. As with small networks, there may be compensatory mechanisms that respond to this flux via network reconfiguration at larger scales. Secondly, larger networks in neural systems may be required to operate in fluid ways based on available resources and depending on other demands on the system. Such situations arise when multiple tasks that recruit overlapping neural regions are

undertaken simultaneously.[27] This creates an additional challenge for stability of function in large networks. To the extent that this challenge is met, large networks will exhibit robustness not just over variations in their usual component parts and processes, but also over which parts and processes are recruited at a particular time for a particular task.

There is evidence that this sort of robustness indeed exists in larger networks. A wide range of studies of neuroplasticity have demonstrated the brain's ability to recover function in the face of injury. The data showing this come from either lesion studies in model organisms or studies of recovery in human patients after traumatic brain injuries.[28] Computational studies of plasticity at the neuron level provide a plausible mechanisms by which such robustness may be achieved (Abbott and Nelson 2000, Albensi 2001, Noppeney et al. 2004). And perhaps more saliently, large scale network analyses comport with the idea that the functions of these networks are highly robust (Bullmore and Sporns 2012, Stromatias et al. 2015). This all serves to suggest that robustness is a ubiquitous feature of neural systems that should be expected to be found across many levels of organization. Insofar as the example of robustness in Purkinje cell burst firing holds implications for causal inference, those implications should be expected to hold, mutatis mutandis, for causal inference throughout neuroscience.[29]

[27] There is ongoing, substantive debate on whether and to what extent the brain accommodates this kind of multitasking. See Fischer and Plessow (2015) for a recent review.
[28] For relevant reviews, see Kolb and Gibb 1999, Kolb and Gibb 2008, Bach-y-Rita 2003.
[29] This is particularly relevant in fMRI research, where a number of causal modeling techniques have been employed in the service of recovering functional connectivity in resting state and block designs—see Henry and Gates 2017 for an excellent review.

3.5 The Significance of Failures of Faithfulness

The discussion in §II concluded that violations of the CFC should have zero probability of occurring in systems in which the parameters governing the causal relationships between variables vary continuously both individually and jointly (conditional on any subset of the others).[30] I have just argued in §III that systems that exhibit robustness will tend to violate the CFC. The natural conclusion would thus be that systems that exhibit robustness have causal relationships that are governed by parameters that do not vary continuously both individually and jointly. In what follows, I'll show that this is plausibly the case with the Purkinje cell example. However, this only tells part of the story. My claim in §III was not merely that examples of robustness provide cases where violations of the CFC are not probability zero, but actually cases where violations of the CFC will be likely. I argue that a similar feature—parameter coupling—drives both failures of joint continuity and the likelihood of violations of the CFC in the Purkinje cell case. I then conclude by considering the implications of the likelihood of violations of the CFC for causal inference in systems that exhibit robustness.

Consider the relevant parameters in the Purkinje cell case, as depicted in *Figure 5*. In that simplified model, postspike inward current is a function of both sodium conductance and potassium conductance, and potassium conductance is also a function of sodium conductance. If we ignore error terms and assume these relationships are linear,[31] we get the following causal model:

---

[30] In the sense that the marginal distributions of each variable should be continuous and the joint probability density function should be continuous.

[31] Linearity is almost certainly an unrealistic assumption, but SGS's measure zero proof is restricted to linear causal models. I assume linearity here both for simplicity and in order to allow this example to make contact with SGS's proof.

K^+ *conductance = b (*Na^+ *conductance)*

Δ Postspike current = a (Na^+ *conductance) - c (*K^+ *conductance)*

Conceptually, the relevant parameters refer to: (a) the effect of sodium conductance on postspike inward current, (b) the effect of sodium conductance on potassium conductance, and (-c) the effect of potassium conductance on postspike inward current. As alluded to in my discussion of the example in §III, the feature that seems to be driving robustness is a sort of parameter coupling. Put simply, it seems that these parameters do not vary independently. Instead they are tied together in a way that ensures stability in postspike current over substantial variation in sodium conductance. We can see this more precisely by spelling out the relationships between these parameters in light of Swensen and Bean's analysis.

As sodium conductance varies, so does the action potential waveform. Potassium conductance, in turn, depends on the action potential waveform. Recall that potassium has an inhibitory effect on postspike conductance (and thus takes negative values). In order for (a) the effect of sodium conductance on postspike inward current to be offset by corresponding change in the effect of potassium current on postspike inward current, (b) the effect of sodium conductance on potassium conductance must vary in proportion to (a). And then to ensure robustness, that change in potassium conductance (due to changes in the action potential waveform) must also be inversely proportional to (-c) the effect of potassium conductance on postspike inward current. To see why this inverse proportionality holds, note that the stronger the effect of potassium conductance on postspike inward current, the weaker the increase in potassium conductance (due to changes in the action potential waveform) needs to be to offset any effect on postspike current due to changes in sodium conductance. Put formally, the relationships between these parameters can be expressed as $b = -a/c$.

Given this analysis, it is straightforward to show that this is a case in which the joint probability density function for these parameters falls to be continuous. First, we can associate each of these parameters with a random variable—respectively, V_a, V_b, and V_c. Recall that joint continuity requires not only that the marginal distribution of each random variable be continuous, but also that each be continuous conditional on any subset of the others. It is reasonable to suppose that V_a, V_b, and V_c are all marginally continuous. However, from the above analysis, we have $V_b = V_a / V_c$. To the extent that this relationship holds, V_b is not continuous conditional on V_a and V_c (to the contrary, it's value would be fixed).

Another way of getting at this connects directly to the measure zero argument. For V_a, V_b, and V_c to be jointly continuous is for there to be a joint probability density function defined over these variables. Such a joint probability density function would specify a four dimensional "curve" over $\mathbb{R}^3$ and would thus assign positive probability only over volumes specified by the products of ranges for each Va, Vb, and Vc. However, again, if Vb = Va / Vc no such probability density function can exist. Note that once we set two of the values, the third is fixed. Thus, the set {(Va, Vb, Vc) $\in \mathbb{R}^3$ | Vb = Va / Vc} has zero volume (it instead specifies an area), and yet receives unit probability. As a result, there is no joint probability density function for Va, Vb, and Vc, but instead just a joint cumulative distribution function that specifies a three-dimensional solid over areas in the above set. This, of course, entails that a measure zero subset of $\mathbb{R}^3$ receives nonzero probability.

The astute reader will have noticed that the equation, $b = -a/c$, is just a transposition of the relationship that must obtain between the parameters in the causal model above for that model to violate faithfulness. So, it is unsurprising, given that this relationship holds, that the random variables associated with the parameters would fail to be jointly continuous, and that the joint distribution function would assign positive probability to measure zero subsets of $\mathbb{R}^3$. These

conditions just follow once a violation of faithfulness is established. But this does not mean the discussion is question begging. Rather it serves to make explicit the consequences of the parameter coupling that seems to be at play in the Purkinje cell case. And that parameter coupling is not merely stipulated but is also backed by a causal explication of the ways the relevant parameters interact.

Of course, it is one thing to argue that this parameter coupling is what drives robustness in the Purkinje cell case, it is another to say that this is generally the way that robustness is achieved in complex systems. At the end of §III, I argued that robustness is widespread in neural systems, and is likely to arise in some form at all levels of analysis. The mechanisms that enable robustness may not always take the form of parameter coupling, though in all cases robustness is likely to cause problems for the CFC. The reason is that the key feature of these complex systems that drives robustness is the fact that systems have to maintain stable functions on timescales that exceed the lifespan of the component parts and processes that support those functions. This in turn requires the system to develop mechanisms that tune parameters such that functions persist despite significance variance in the relevant components parts and processes that support those functions.

So, what implications for causal inference should be gleaned from the likelihood of failures of faithfulness in systems that exhibit robustness? Recall from §I that the CFC plays key roles in causal hypothesis testing and causal discovery. With respect to causal hypothesis testing, the CFC allows inference from a causal link between two variables to the claim that those two variables are not probabilistically independent (given appropriate conditionalizations). So, if you start with a proposed causal link, but then discover that the linked variables are in fact probabilistically independent, the CFC dictates that you then reject the causal link.

The hypothesis that motivates Swensen and Bean's study in fact exhibits this inference pattern but arrives at a different conclusion. A previous study (Swensen and Bean 2003) had revealed that bursting in Purkinje cells occurs as a result of multiple large inward and outward currents combining to produce a small net inward current. They state their hypothesis for the 2005 study in reference to these previous results, "[t]hese results suggest a fine balance of postspike currents in which a small change in the size of any individual current, through slow inactivation, modulation, or other perturbation, could have dramatic effects on bursting" (Swensen and Bean 2005, p.3509). In other words, given that previous results have demonstrated a causal link between certain postspike currents (particularly, TTX-sensitive sodium currents) and burst firing, one would expect strong probabilistic dependence between those postspike currents and bursting. Of course, their study found precisely the opposite. If we accept the CFC, then the appropriate response to this result would be to deny the presence of a causal link; an alternative response would be to reject the CFC. Swensen and Bean essentially opt for the latter.

Compare this to the TARP study discussed in §II. Rossi et al.'s (1980) inference pattern is essentially identical to Swensen and Bean's. They considered a potential causal link between monetary disbursements and recidivism and also found that the two were probabilistically independent. Rather than rejecting the causal link, Rossi et al. instead rejected the CFC. But in the case of the TARP study, this seems to have been the wrong conclusion. The difference between these two cases mirrors the difference between the TARP study and the gravity example. The causal link between monetary disbursements and recidivism is, at the start, on much shakier ground than the causal link between TTX-sensitive sodium currents and bursting. Further, there is no reason to expect the sort of parameter coupling that supports robustness in the Purkinje cell case would obtain between monetary disbursements and recidivism. We can conclude that the CFC is

a useful and intuitive principle for causal hypothesis *formation*; however, with respect to causal hypothesis *testing*, the CFC should not be taken as a steadfast rule. Whether a probabilistic independence counts as evidence of robustness or evidence of causal independence ultimately depends on the strength of extraneous evidence for the causal link.

This last point bears on causal discovery, in a certain sense. To the extent that some probabilistic independence is discovered between variables for which there is strong extraneous evidence of a causal link, this independence serves as evidence of robustness (and hence a violation of the CFC). The discovery of that probabilistic independence thus encourages investigation into the means by which robustness is achieved. This is effectively the structure of the reasoning in Swensen and Bean's study. However, this point is more an offshoot of the implications of failures of the CFC for causal hypothesis testing than a point of relevance to causal discovery in the sense most relevant to DAG causal modeling, as discussed in §I. In the DAG framework, causal discovery consists in inferences from probabilistic data to causal relationships.

To that end, recall that the CFC licenses inference from probabilistic independence between two variables (given appropriate conditionalizations) to the absence of a causal link between those variables. Causal discovery inferences are most relevant to contexts where causal relationships are unknown. Thus, on one hand, the causal discovery implications of the CFC are less relevant in the Purkinje cell example simply because the causal relevance relationships in the system were already fairly well understood. Nonetheless we can see, through this example, the basic shape of the challenge robustness poses to causal discovery. Suppose we had no prior knowledge of the causal factors driving burst firing in Purkinje cells. If we had only a joint probability distribution showing the probabilistic relationships between the various conductances and the size of the net interspike currents driving burst firing, we would be at a loss to infer any causal links.

This issue has the potential to muddy the waters of causal discovery in serious ways. For instance, suppose we are interested in exploring causal links between particular genetic abnormalities associated with depression—e.g. allelic variation in serotonin transporter gene polymorphism (Haenisch et al. 2013)—and neural abnormalities also associated with depression—e.g. decreased amygdala volume in unmedicated depression (Hamilton et al. 2008). The typical way to search for such links would be to perform a genome-wide association study (GWAS) among individuals suffering from depression. If these association studies were to show that the two factors are probabilistically independent, what would be the appropriate conclusion to draw?[32] Should we infer that this is a case, similar to the TARP study, where the CFC holds and so the absence of probabilistic dependence implies there is no causal link? Or should we assume that this is a case, similar to the Purkinje cell example, where the brains of depressed individuals exhibit compensatory pathways that produce decreased amygdala volume over a range of variation in relevant genetic factors? In the absence of some supplementary information about the causal structure of these systems, it is difficult to see any way to offer principled answers to these questions.

There are a number of conclusions we can draw from this. First, independent of its likelihood of holding in any particular domain, the CFC is a useful guide for causal hypothesis formation. For any proposed causal link between two variables, a natural way of testing that link is to analyze the dependence between those variables (given appropriate controls). A negative result (i.e. the discovery of probabilistic independence between the variables) can be either reason to reject the causal link or reason to look for counterbalancing causal pathways. Which conclusion to draw

[32] In fact, associations between allelic variation in serotonin transporter polymorphism and amygdala volume have been inconsistent, with some studies reporting associations and other, more powerful studies failing to replicate those associations (Pezawas et al. 2005, Scherk et al. 2009, Stjepanovic et al. 2013).

depends, primarily, on the strength of the independent evidence supporting the proposed causal link. Second and relatedly, robustness poses serious challenges to causal discovery. To the extent that robustness is likely to occur likely in a particular domain, it is unclear that any causal information can be gleaned from probabilistic independence. Again, that independence can reflect either the absence of a causal relationship or the presence of counterbalanced causal pathways. No determination between those interpretations can be made in the absence of supplementary knowledge of the causal structure of the system. But requiring such supplementary causal knowledge undermines the purpose of causal discovery, in the sense at issue in DAG causal modeling.[33]

3.6 Conclusion

In this chapter, I have argued that systems that exhibit robustness will tend to violate causal faithfulness. I offered detailed analysis of the example of robustness of burst firing in Purkinje cells, and showed how this example demonstrates a violation of faithfulness. I argued that the key feature driving robustness in this case is a form of parameter coupling that is well characterized in the causal dynamics of the system. This parameter coupling demonstrates how failures of the CFC can not only fail to be probability zero but can also be highly likely.[34] I argued further that robustness is likely to be found in complex systems that maintain stable functions across timescales that exceed the lifespan of the component parts and processes that support those functions. I

[33] Hat tip to acknowledge that this conclusion is in keeping with Nancy Cartwright's (e.g. 1999a) slogan: "No models in, no causes out."
[34] "High likely" to the extent that it makes sense for there to be a probability distribution over the possible parameter values.

concluded by arguing that this likelihood of failures of faithfulness has significant consequences for both causal hypothesis testing and causal discovery.

4.0 Robustness, Modularity, and Cyclicity

4.1 Introduction

In this chapter, I continue to explore the epistemic consequences of functional robustness (qua an instance of causal explanatory multiple realization). Mitchell (2008, 2009) has argued that systems that exhibit robustness fail to be modular, in the sense of having components that are independently disruptable. This is significant because modularity is a crucial feature of difference-making accounts of causal inference, particularly interventionism (Woodward 2003). If causation is to be understood in terms of difference-making, one must be able to manipulate causal variables in a system without changing other causal relations in that system. Mitchell argues that this requirement fails when robustness is achieved through some form of reorganization of causal structure in response to localized experimental manipulations. Her support for this conclusion draws from evidence of genetic robustness—specifically, experiments that show that disruption to individual genes often has no phenotypic effect in organisms. She concludes from this that interventionist theories of causal inference are not viable in complex systems, and that instead new methods of causal inference and theories of explanation are needed.

I argue, by contrast, that closer inspection reveals that instances of robustness are often indicative of feedback loops driving systems toward particular outcomes. That is, robustness does not indicate a failure of modularity, but instead a failure of acyclicity. Causal inference in cyclic systems presents its own set of challenges, but those challenges do not support general skepticism of difference-making accounts of causation or support a call for radically different methods of

causal investigation. Indeed, I show that modularity is a crucial component of unearthing the causal structure of cyclic systems.

In §I, I provide general background on interventionist accounts of causation and discuss the role of modularity in those accounts. In §II, I consider some of the standard objections to modularity and offer a slight amendment to the concept. In §III, I reconstruct Mitchell's argument that functional robustness in genetic knockout experiments is incompatible with modularity. I argue to the contrary that modularity plays a crucial methodological role in knockout experimentation. In §IV, I argue that Mitchell mislocates the challenge functional robustness poses to theories of causation and methods of causal investigation. Rather than undermining modularity, I argue that functional robustness reveals the presence of feedback control—i.e. the presence of cyclical causal structure. I show how modularity can again play a crucial role in unearthing the cyclic causal structure. I conclude by noting some of the challenges cyclicity poses to theories of causation and causal inference.

4.2 Modularity and Interventionism

The interventionist theory of causation developed from the same roots as the graphical causal modeling (GCM) framework discussed in chapter three. It can be thought of as the semantics of the inferential methods and principles associated with that framework.[35] Interventionism interprets the meanings of causal claims as claims about how the world would be different given certain

[35] See Woodward 2008 (pp.200-201) for discussion of interventionism as a semantic project. For the general framework of interventionism and explicit discussion of its relation to graphical causal modeling, see Woodward 2003.

changes, where those changes are understood as hypothetical idealized experimental manipulations. So, to say X causes Y is just to say that, if all other relevant features are held constant, changing the value of X will result in as change in the value of Y. Thus, to take an example from the previous chapter, to say that alcohol abuse causes psoriasis just is to say that, given appropriate controls (or conditionalizing on the other relevant variables), adjusting alcohol consumption will adjust risk of developing psoriasis. Note that the claim is not that it must be possible to actually perform this intervention. For instance, we can understand the claim that the moon's gravity causes oceanic tides as a counterfactual claim about how the tides would be affected if the moon didn't exist or were closer or further from Earth without actually being able to perform those interventions.

Interventionism has been an influential part of the shift away from the deductive-nomological model toward causal theories of explanation, as discussed in the first chapter. Like other causal theories of explanation, interventionism takes explanation to consist in illuminating causal relationships that give rise to a phenomenon, rather than subsuming that phenomenon under laws of nature. In addition, interventionism replaces the concept of universal laws of nature with the concept of *invariant* causal generalizations. To say that a generalization is merely invariant, as opposed to universal, is to allow that it may hold only in certain circumstances, e.g. given certain background conditions or restricted ranges of parameter values. This notion of invariance is a significant improvement over the previous conception of laws qua universal generalizations, as it allows us to nonetheless use and make sense of generalizations that hold only in particular sets of circumstances.

Take, for instance, the snap mousetrap, briefly discussed in chapter one and depicted below. Our understanding of the causal operation of snap mousetraps is roughly as follows. The trap is

set by pulling the hammer, against the force of the spring, in an arch over the platform; the holding bar is then placed over the hammer and held by the catch on the opposite side of the platform. When the catch is tripped, the holding bar is released and the potential energy of the spring's force again the hammer is converted to kinetic energy, causing the hammer to slam onto the opposite side of the platform. This causal story does not hold universally—instead it is invariant over a range of values that the relevant parameters that determine the functioning of the system can take. For instance, if the spring is too strong relative to the sensitivity of the catch or the strength of the holding bar, the system will cease to function in accord with this causal chain.

A different form of stability, the notion of *modularity,* is also a key component of the interventionist framework.[36] Modularity refers to the independent manipulability of causal relationships in a system. Independent manipulability here means that changing one functional relationship within the system does not cause changes to other functional relationships within the system. Woodward (2008) defines modularity as follows:

> Modularity: A system of equations is modular iff (i) each equation is invariant under some range of interventions on its independent variables and (ii) for each equation, it is possible to intervene on the dependent variable in that equation in such a way that only the equation in which that dependent variable occurs is disrupted while the other equations in the system are left unchanged. (Woodward 2008, p.221)

This concept of modularity is an important, though controversial, component of the interventionist framework.[37] If causal claims are understood as counterfactual claims about how the world would be different under different hypothetical manipulations, it is important that

---

[36] We encountered this concept in passing in chapter three in the form of the notion of autonomy from Aldrich (1989), and as incorporated by Pearl (1999).

[37] In less cautious moments, interventionists have described modularity as intrinsic to the concept of causality (e.g. Hausman and Woodward 1999, p.550), and have attempted to use modularity in the service of deriving the causal Markov condition (Hausman and Woodward 1999, Hausman and Woodward 2004). See Cartwright 2002 and Steel 2006 for objections to the latter efforts.

systems exhibit a certain amount of stability in the face of those hypothetical manipulations. Particularly, one must be able to manipulate causal relationships in a system without changing other causal relations in that system, otherwise there would be no way to attribute particular causal content to the contribution of that causal relationship.

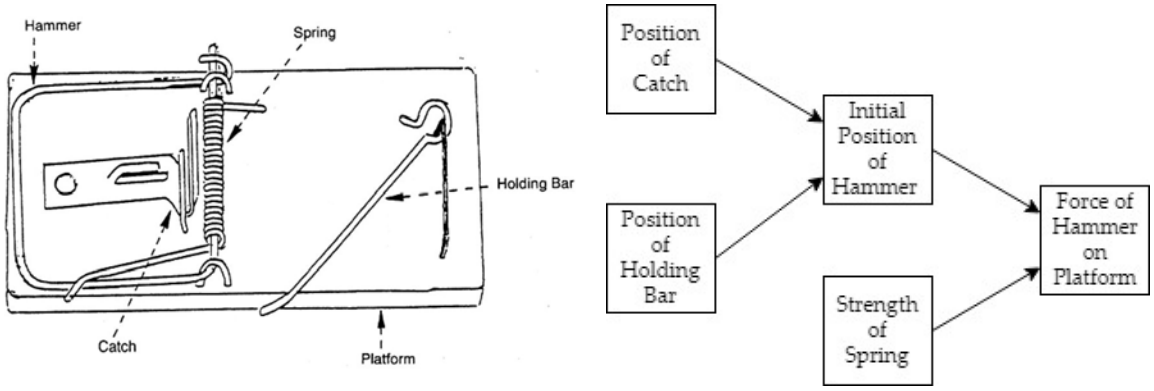


Figure 6: Illustration of a snap mousetrap with a causal diagram.

Note: the causal diagram depicts the relevant factors bearing on the force exerted by the hammer on the platform upon release of the catch.

Consider, again, the snap mousetrap, represented above in *Figure 6*. This system is modular in the relevant sense. For instance, suppose we are interested in understanding the relevant causes that determine the force that the hammer exerts on the platform, as represented in the associated causal diagram. That force is a function of the initial position of the hammer (when the trap is set) and the strength of the spring. The initial position of the hammer is a function of the positions of both the catch and the holding bar, where position is understood to contain information about both the height of the connection point and the relative distance between each respective connection point and the top of the crossbar of the hammer. Intervening of one of these variables, say, the strength of the spring, has no influence on the other causal relationships in the system. As long as we are within the values of parameters and variables that allow the system to function, which is the point of (i) in Woodward's definition, we can alter the strength of the spring without causing

changes that cascade through the rest of the system. As a result, if we want to determine the causal relationship between the strength of the spring and the force exerted by the hammer on the platform when the catch is tripped, we can simply vary the strength of the spring and observe the resulting change in force.

Consider, by contrast, what it would mean for modularity to fail in this system. Suppose our only means of manipulating the strength of the spring involved adjusting the tightness of the coils, and that this, in turn, affected the position of the catch (say, in order to clear the adjusted coils). To the extent that we grant this, this system would violate modularity. Manipulating the strength of the spring would alter the initial position of the hammer, and in the process would confound the effect of changes in strength of the spring on the force of the hammer on the platform. The consequence is that it would be impossible to isolate the causal contribution of the strength of the spring on the force exerted by the hammer on the platform.

4.3 Challenges and a Modification

Nancy Cartwright has been a staunch skeptic of modularity (see, e.g., Cartwright 2001, 2002, 2007). Her opposition to the concept has generally been advanced through apparent counterexamples, focused on everyday items and their internal mechanics. The idea is perhaps that if modularity fails for such mundane, everyday items, it shouldn't be expected to hold in more complex systems. One of Cartwright's recurrent counterexample involves a common toaster. She explains as follows,

> The expansion of the sensor due to the heat produces a contact between the trip plate and the sensor. This completes the circuit, allowing the solenoid to attract the catch, which releases the lever. The lever moves forward and pushes the toast rack

> open… I would say that the movement of the lever causes the movement of the rack. It also causes a break in the circuit. Where then is the special cause that affects only the movement of the rack? Indeed where is there space for it? The rack is bolted to the lever. The rack must move exactly as the lever dictates. So long as the toaster stays intact and operates as it is supposed to, the movement of the rack must be fixed by the movement of the lever to which it is bolted. (Cartwright 2002, pp.70-72)

The toaster fails to be modular, because the mechanism that pops the toast up is part of the same system that completes the circuit and delivers electrical current to the heating elements. It makes sense to turn the heating elements off when the toast is done, and it makes sense to pop the toast up so you can retrieve it without burning your hand, and there's some benefit to those things to happening at the same time, so toaster designers just connect the two mechanisms. The system fails to be modular in the sense that there are not independent causal switches that can be intervened upon to isolate those separate causal mechanisms.

This is not a particularly compelling counterexample. As Woodward (2008) and Steel (2010) have pointed out, it would be easy to alter the system in a way that would enable separate manipulability of the rack and circuit. For instance, if the toaster were designed like, say, a toaster oven, where the movement of the rack (horizontally, rather than vertically) is instead tied to the opening of the door and is isolated from the circuit delivering electricity to the heating elements. Note, however, that with a toaster oven, unlike a toaster, the means for adjusting temperature is not separable from the on/off switch for the circuit (there is usually just one dial); whereas with a toaster, the on/off switch, which is controlled by the lever, is separate from the browning dial that controls either the temperature or the time of toasting. At any rate, the point is that plenty of things are designed in a way that connects distinct causal mechanisms but should not really be what's at issue in the notion of modularity. It would be no great feat of engineering to separate these mechanisms in everyday items like toasters—e.g. by unbolting the lever from the rack.

Cartwright's response, contained in the quote above, would presumably be that altering the toaster in this way should be prohibited, or changes the toaster in substantial enough ways that it no longer continues to operate as a toaster. But this will not do. It is antithetical to how we reason about causal systems. Isolating components and tinkering with them to see what their individual effects can contribute to the systems that contain them is fundamental to most areas of engineering as well as causal investigation of complex systems.[38] For instance, the vast majority of experiments in cellular neuroscience involve isolating individual neurons from nervous systems and tinkering with them in vitro. Of course, there is always some worry that such isolation will generate causal claims that fail to generalize back to the system of interest. That is, there is always concern that causal claims won't generalize from the lab back to the world. But if this were generally the case, modern investigations of complex systems would be completely ineffective.

Though not a compelling counterexample, the toaster case is instructive in that it encourages more precision in what it means for a system to be modular. Modularity is typically described as a property of representations, or systems of equations (see, e.g., Woodward 2003, p.48; Woodward 2008, p.221). There is a perfectly clear sense of what modularity means in such contexts, as is specified in the definition from Woodward (2008) in the previous section. However, it is perhaps more useful to think of modularity in the first place not as a feature of representations, but as an assumption about causal systems that licenses certain inferences and encourages particular ways of conducting causal investigations. This assumption amounts to something like the following:

> Modularity*: to say a causal system is modular is to imply that each cause in the system provides an isolated causal contribution to its respective effect.

[38] Woodward calls this a "Galilean" approach to the function of experiments (2008, p.229).

The notion of isolation here can be taken to mean independence of the causal relationships in the system. Something akin to modularity* is an implicit methodological principle that drives the sorts of causal investigation just mentioned that involve deconstructing systems and tinkering with their components in relative isolation to determine their particular causal roles within some system. Reframing modularity in this way captures the same core feature at issue in Woodward's definition but avoids some of the problems that definitions invites.

The key differences between Woodward's definition of modularity and modularity* are: (1) modularity* stipulates that modularity is a property of causal systems rather than a property of representations, and (2) modularity* stipulates only that causal contributions of individual causes are isolated, not that they are *isolable*. With respect to (1), the formal definitions of modularity invariably take the form of claims about systems of equations. However, elsewhere, Woodward (2003, 2008) and Hausman and Woodward (1999, 2004) vacillate on whether modularity should be construed as a property of representations or a property assumed of causal systems in the world. For instance, Woodward writes, "[i]t is natural to suppose that if a system of equations correctly and fully represents the causal structure of some system, then those equations should be modular" (Woodward 2008, p.48). In other words, Woodward here seems to regard modularity as a feature of the world and so a criterion of correctness for causal representations of the world. It is also common for descriptions of modularity to get tied to the notion of distinct causal mechanisms (in the world).[39] At any rate, the key point of (1), thinking of modularity as a property of the world rather than a property of representations, is that it creates room for the distinction at play in (2).

[39] See, e.g., Hausman and Woodward (1999), p.549, where modularity is offered as a criterion on mechanism individuation.

With respect to (2), the standard definition of modularity requires that it be possible to perform an intervention that isolates the causal contribution of each cause (dependent variable) within a system. This encourages thinking of modular systems as those in which there are particular switches or dials for each relevant variable within the system and that each switch or dial must be able to be adjusted in a way that is completely independent of the others. Moving away from the requirement of isolability cuts off this temptation. The fact that a causal contribution is presumed to be isolated from the other causal relationships in a system does not entail that it is isolable by means of some particular intervention. That is, even if causal contributions are not isolable by means of particular interventions, there may be nonetheless ways to infer isolated causal contributions.

Take, for instance, our imagined "modularity violating" snap mousetrap from the end of the previous section. The idea there was that manipulating the strength of the spring had the effect of also changing the position of the catch. In such a system, there fails to be a single intervention that isolates the causal contribution of the strength of the spring from the effect of the position of the catch. However, as long as the corresponding change in position of the catch is regular and quantifiable, it would be a fairly trivial matter of geometry to figure out how the initial position of the hammer covaries with the strength of the spring. The effect of that change in initial position on the force of the hammer can then be calculated and subtracted off the overall effect of manipulating the strength of the spring in order to isolate the causal contribution of the change in the strength of the spring from the change in position of the catch.

This distinction between isolated and isolable causal contributions bears on the point that typical objections to modularity, like Cartwright's toaster, tend to focus on. Cartwright takes the toaster to be a counterexample precisely because the design of the system precludes independent

manipulation of the rack and electrical circuit without reengineering the system. But if modularity only implies that there are isolated causal contributions in the target system, then the fact that there are not easily isolable is beside the point. Moreover, if we assume that the system is in fact modular, in the sense of modularity*, then the natural way to home in on the isolated causal contributions of each component is to deconstruct/reengineer the system in order to separate the mechanisms responsible for different functions—which, in this case, involves separating the mechanism for opening the rack from mechanism delivering electricity to the heating elements. The fact that the same physical component plays a role in two separate causal chains in no way precludes that component's ability to make an isolated contribution to each chain.

The distinction is also relevant when available technology limits the precision of experimental manipulations. Consider Karl Lashley's (1960) efforts to isolate the function of visual cortex in blinded rats' abilities to learn and navigate mazes, as discussed by Bogen (2004, p.19-24). The rats were trained to navigate a maze with vision intact, then blinded and (re)trained to navigate the maze, and then their visual cortices were removed and their abilities to navigate the maze were tested. The ideal means of testing this would have been to ablate the rats' visual cortices without affecting any surrounding regions. However, given technology of the time, such precise localized ablation was not possible, other surrounding areas inevitably got damaged as well. Bogen explains Lashley's solution,

> To work around this he lesioned the visual cortex in a variety of different ways, each one of which was designed to do collateral damage to a different adjacent structure. In one group of rats, the hippocampal lobes were lesioned along with the visual cortex, and the auditory cortex was spared. In another group, the auditory cortex was damaged and the hippocampal lobes were spared. And so on for each of the other regions next to the visual cortex. In a final group Lashley lesioned the visual cortex while sparring as much of the rest of the brain as he could. (Bogen 2004, p.20)

In other words, Lashley exerted control over which surrounding areas were affected and varied them systematically. He set a number of different conditions in all of which visual cortex was lesioned, but different adjacent regions were lesioned in each condition. Proceeding in this way, Lashley then tested the rats' performance at navigating a maze after lesioning. Though the rats had been retrained on the maze after being blinded, their performance after lesioning was uniformly impaired on the maze task (as much or more so than the impairment observed after blinding). Lashley concluded that visual cortex is involved in rats' abilities to learn and navigate mazes (in addition to its role in processing visual information.

What is clear from this description of Lashley's experiment is that he was operating with an assumption like modularity*. In particular, he assumed that visual cortex was playing an isolated role in maze navigation in blinded mice; this is what motivates his systematically varied "fat-handed" interventions. However, that role was not isolable given the lack of precision in experimental techniques available at the time. This, arguably, causes issues for Woodward's definition of modularity, which is part of Bogen's purpose in bringing up the case.[40] Those issues are again sidestepped with the tweaked notion of modularity*, which also serves to illuminate the motivation behind Lashley's approach.

[40] Bogen cites this example in the service of an argument against interventionism (and counterfactual theories of causation, more generally). Bogen argues of this case that Lashley made no reference to the potential ideal intervention (in Bogen's terms "immaculate manipulation of a system which meets the modularity requirement") that he was clearly trying to approximate, and no counterfactual claim about what would have resulted had that ideal intervention been performed. Woodward 2008 counters Bogen's interpretation by arguing that Lashley was trying to understand what would happen if he were to perform an ideal intervention even though, as a practical matter, he could not perform such an intervention. Thus, Woodward argues, the causal claims Lashley concluded regarding the role of visual cortex in learning and navigating mazes are well characterized in the interventionist framework. See Woodward 2008, pp.209-211, for further discussion.

4.4 Mitchell's Challenge: Robustness and Modularity

Sandra Mitchell (2008, 2009) has challenged modularity from a different angle. She argues, rather than simple systems like toasters, complex systems, particularly those found in biology, often fail to be modular. Mitchell's argument is based on anomalous gene knockout experiments, which she argues indicate a form of genetic robustness very similar to the examples of neural robustness that have been considered throughout this dissertation. She argues that genetic robustness provides instances in which intervening on some particular causal relationship reconfigures other causal relationships in order to maintain the function in question in the face of the intervention. She concludes from this that genetic systems violate modularity and thus that new theories of causation are needed to account for the complex causation found in biological systems (see, in particular, Mitchell 2009, Ch4).

To understand Mitchell's argument, it is first necessary to provide some background. Genetic knockout is a targeted form of intervention that involves inserting artificial, nonfunctional DNA into the chromosomes of embryonic stem cells in vitro. The embryos are then transplanted into a female uterus of some particular model organism, most typically a mouse, and allowed to develop. The resulting mouse pups are heterozygous knockouts, which can then be bred to create homozygous knockouts. These techniques provide a precisely targeted means of manipulating genes to determine their roles in supporting different phenotypes (cf. the discussion of Lashley's experiment on rat visual cortex discussed at the end of the previous section). One might thus expect knockout experiments to be exemplars of interventionist accounts of causality—precisely controlled manipulations of dependent variables (genes) lead to changes in independent variables (phenotypes), and a causal link is inferred between those variables. And indeed, many knockout experiments proceed in just this way. Schofield et al. (2012) note in a review that of the ~25,000

mouse genes with protein sequence data, ~8,200 have identified phenotypes.[41] That is a significant amount of success in establishing causal links between genes and phenotypes, especially when considering that knockout techniques had only been in practice for 20 years when their review was published.

However, not all genetic knockouts have associated phenotypic changes. Roughly 15 percent of gene knockouts are developmentally lethal. Aside from those, and to the point crucial for Mitchell's argument, some sizable percentage of knockout experiments show little to no evidence of phenotypic change. These are the so-called "anomalous" knockout experiments, and Mitchell cites their proportion at roughly 30% of all knockout experiments (Mitchell 2008, p.700, Mitchell 2009, p.68). Other estimates fall in the 10-15% range (Barbaric et al. 2007), but the accuracy of this number is actually quite hard to gauge, for reasons that will be explored further below. Nonetheless, in some subset of these cases, Mitchell argues that systems exhibit a form of robustness similar to those I've considered in earlier chapters in the context of neural systems. Specifically, she homes in on cases where the robustness of phenotype to genetic variation may be due to reorganization of causal structure.

Mitchell's argument, offered primarily in her own words, proceeds as follows,

> [I]n up to 30% of double knockouts there is little or no evident phenotypic consequence of knocking out a gene. The cases where the knockout produces no substantive phenotypic difference may point to the dynamical plasticity of the genetic network. Robustness due to redundancy or degeneracy will make it difficult to make inferences about the normal causal structure from an intervention or perturbation of the system. (Mitchell 2008, p.700)

[41] See also the massive databases available online collating information on the mouse genome (www.informatics.jax.org) and phenome (phenome.jax.org).

Recall that we encountered this distinction between redundancy and degeneracy (what I've called functional robustness) in the first chapter. Redundancy simply involves a back up copy of a gene substituting for it in the knockout case. When robustness is achieved through redundancy, the other causal relationships within the system need not change. Thus, these cases do not challenge modularity. She continues her reasoning in accord,

> The absence of a phenotypic change even when all redundant copies of a single genetic component are knocked out could indicate that the network itself has reorganized to compensate for the loss of the gene. If so, parts of the network that in the normal state would be described by one set of functional relationships *change* their interactions in response to the experimental intervention to produce a product similar to that of the unperturbed system. (Mitchell 2009, p.71)
>
> It appears that a degenerate or robust system where a genetic network reorganizes when some piece of it is knocked out is not independently disruptable. That is, one gene in the network functions as a causal contribution to the phenotypic effect under normal internal conditions, but functions in a different way when another part of the network is removed… Thus Woodward's condition of modularity is not met. (Mitchell 2009, p.77)

Thus, Mitchell argues that degeneracy (i.e. functional robustness) provides the best explanation for the results of some anomalous knockout experiments, especially when redundancy can be ruled out. And it is these instances of robustness that undermine Woodward's notion of modularity. Her conclusions from here are sweeping: that new concepts of causation are necessary and no uniform methodological prescriptions (e.g. regarding experimentation or how to conduct causal investigation) will be adequate for such complex causal systems.

While I am generally sympathetic to Mitchell's project (as should be no surprise given the topic of this dissertation), this argument moves too quickly to support such sweeping conclusions. Further, with respect to the more measured conclusion regarding the prospects for modularity in systems that exhibit robustness, I believe that Mitchell mislocates the challenge that robustness

poses to theories of causation and causal inference. Spelling this out requires first taking a step back.

Results from anomalous knockout experiments are not straightforwardly attributable to robustness. Mitchell is well aware of this. In the setup for her argument (2009, p.68), she cites a famous quote from Mario Capecchi, who received the Nobel prize in 2007 for his research in the field: "I don't believe in complete redundancy. If we knock out a gene and don't see something, we're not looking correctly." She also offers a quote from the opposite end of the ideological spectrum from Robert Weinberg, a pioneer of research on the genetics of cancer: "The big surprise to date is that so many individual genes, each of which has been thought important, have been found to be nonessential for development."[42] Other researchers (e.g. Greenspan 2001, Edelman and Gally 2001), with whom Mitchell sides (2009, p.68), offer a distinct possibility—that phenotypes may be impervious to gene variation as a result of robustness in genetic networks.[43] There are thus three broad possibilities that must be considered in the interpretation of anomalous knockout experiments: (1) there is actually a phenotypic difference that simply has not yet been discovered, (2) the DNA is nonessential, (3) the system exhibits robustness, either in the form of redundancy or degeneracy. Interpreting anomalous results to imply (3) is thus not uncontroversial.

I take (2) to be a conclusion of last resort, as it effectively ends inquiry despite the presence of other live options. However, (1) merits further scrutiny. Barbaric et al. (2007) provide an excellent review, extensively detailing possible explanations of anomalous knockout results. They offer the following set of options.

---

[42] Both quotes can be found in Travis 1992.

[43] Greenspan (2001) demonstrates a proof of concept for functional robustness in genetic networks, showing that model genetic networks may be able to maintain functions by reorganizing in the face of deletion of particular genes. Edelman and Gally (2001) introduce the term degeneracy and discuss a number of potential domains in which it may arise.

> If inactivation of a gene does not lead to an observed abnormal phenotype, there are three possibilities: (i) the abnormal phenotype is present under the conditions currently being used but is yet to be discovered, (ii) the abnormal phenotype will only become evident under environmental conditions that have not yet been tested or (iii) there is no abnormal phenotype. (Barbaric et al. 2007, p.92)

The first two options refer to distinct ways phenotypic differences can be obscured. First, (i) there may be a phenotypic difference that is present in experimental conditions but has not yet been discovered. One major methodological issue that creates this possibility is the lack of standardization in phenotyping protocols. Some large-scale efforts have been made to standardize protocols, e.g. the German Mouse Clinic (Gailus-Durner et al. 2005). In its first several years of operation, the German Mouse Clinic analyzed more than 80 knockout lines and discovered previously uncharacterized phenotypes in 95% of those lines (Fuchs et al. 2009). The ongoing discovery of new phenotypes for different mutant strains should give pause to the idea that anomalous results should be taken at face value. However, as protocols become more standardized and more exhaustive, this should become less of an issue. It is difficult to estimate the impact this will have on the proliferation of anomalous knockout results, but it is likely to be significant.

Second, (ii) many phenotypes are only apparent in particular environmental contexts. Thus, it may be that anomalous knockout results are due to experimental conditions that do not provide the environmental conditions necessary for the phenotype to manifest, rather than reflecting an actual lack of phenotypic variation. For instance, Chen et al. (1997) discovered an exocrine gland dysfunction as a result of melanocortin 5 receptor (MC5-R) knockout in an unexpected way. They write,

> No readily visible phenotype was apparent in MC5-RKO mice… Appearance, behavior, growth, muscle mass, adipose mass, reproduction, basal and stress-induced corticosterone, glucose, and insulin levels in these animals were indistinguishable from wild-type littermates. More subtle physiological phenotypes of the knockout were studied by examination of responses to exogenous

> melanocortin peptides in biological assays… None of these assays produced identifiable differences between the wild-type and knockout animals. During a stress-induced analgesia assay in which the mice are made to swim for 3 min to activate the hypothalamic-pituitary-adrenal axis, it was observed that the knockout animals remained wet for a longer period of time than littermate controls. This effect was then identified to result from nearly double the water retention in the coat of the MC5-RKO, resulting in severe thermoregulatory defects in the animal as well. (Chen et al. 1997, p.794)

In other words, MC5-R knockout was previously an anomalous knockout gene, showing no phenotypic variation despite a wide range of phenotyping assays. Fortuitously, during an assay to test for abnormal stress response in which mice are forced to swim, Chen et al. noticed an unrelated abnormality—the coats of the knockout mice stay wet longer than the coats of wild-type mice. It turns out this abnormality reflects an exocrine gland dysfunction due to the MC5-R knockout.

Again, it is difficult to estimate what proportion of anomalous knockout experiments are best accounted for in this way. At any rate, my point in raising this is not to come up with any estimate, but rather to reflect on appropriate methodology. As we've seen, Mitchell argues, on the basis of anomalous knockout results, that new methods of causal inference and new accounts of causal explanation are needed. In particular, she argues that when faced with anomalous knockout results, researchers should not assume modularity, but should instead assume that some more complex causal structure is in play. On one hand, even given the preceding discussion, it is not unreasonable to expect that at least some anomalous knockout experiments are indicative of more complex causal structures. However, it is clearly premature to infer from this that standard forms of experimental design, captured by the interventionist framework, and assumptions of modularity* should be abandoned. Modularity*, as a methodological principle, pushes researchers toward discovery of new phenotypes, rather than accepting anomalous knockout results at face value.

4.5 Answering Mitchell's Challenge: Robustness and Cyclicity

I have just argued that, from a methodological perspective, Mitchell is wrong that modularity* should be abandoned as a principle guiding causal investigation in genetics. However, my discussion does not support the conclusion that all anomalous knockout results are due to undiscovered phenotypes. This thus leaves open the (likely) possibility that some results will be best explained by reference to functional robustness (aka degeneracy). Moreover, I have argued in previous chapters that functional robustness is widespread in neuroscience and has serious implications for causal inference. So my argument that modularity* is the right methodological principle to retain in genetics is beside the point to this main issue: what are the implications of functional robustness for modularity*?

Consider the following, from a discussion of robustness from O'Leary (2018): "If an insect loses a leg, it may or may not lose the ability to walk. But the biomechanical relationships between the remaining legs will be fundamentally altered" (O'Leary 2018, p.182). Autotomy (self-amputation) of appendages, in fact, occurs in many taxa, including vertebrates, echinoderms, crustaceans, and arachnids (Wrinn and Uetz 2007). Spiders are particularly interesting cases because their legs play vital roles not only in locomotion but also as sensory organs—the tiny hairs on spider legs are capable of detecting minuscule vibrations. Yet they can lose two to three legs, and often do as a result of autotomy, and nonetheless retain their abilities to walk and detect prey.

This case gets to the heart of the issue with the relationship between modularity and robustness. On one hand, appendages are generally regarded as exemplars of modularity in biology (e.g. Williams and Nagy 2001). Granted, the notion of modularity in biology is distinct from, though it bears similarity to, modularity* as well as the technical definition of modularity in interventionism. And appendages do seem to satisfy modularity* also; they provide isolated causal contributions to

the capacities they're involved in. However, in certain organisms, like spiders, removal of appendages leads to changes in other causal relationships within the system—particularly, the biomechanical and sensory relationships between the remaining legs. This thus seems to at once satisfy and violate modularity*.

The tension here can be resolved by considering more carefully how the changes in other causal relationships within the system occur. In the immediate aftermath of leg removal, a spider's ability to walk will actually be seriously impaired. This is because the biomechanical relationships between the remaining legs do not adjust automatically. That is, the remaining legs are still controlled by motor patterns predicated on the organism having all eight of its legs. Thus, it is more accurate to say that the biomechanical relationships adjust over time in response to feedback that enables recovery of function. This is significant because it shows that a system can exhibit functional robustness while nonetheless satisfying modularity*. The components of a causal system can provide isolated causal contributions to their effects, which can be assessed on short timescales immediately following an intervention that disrupts one of those causal relationships. And then, on longer timescales, other causal relationships may change as a result of feedback within the system in response to that intervention, and those changes may enable recovery of function—i.e. functional robustness.[44]

The upshot is that in systems where functional robustness is achieved via feedback control, there is a temporal gap between the intervention and the recovery of function. This gap creates

[44] As an aside, this shows another way in which modularity* is preferable to Woodward's definition of modularity. Different intervention techniques operate on different timescales (compare, e.g., pharmacological knockout and genetic knockout in Swensen and Bean 2005). That is, different ways of manipulating dependent variables target different timescales. Modularity* avoids this timescale dependence by removing the requirement of isolability (or possibility of intervention).

room for a productive notion of modularity* to play a role both as a criterion on adequate causal explanations and as a guiding principle in causal investigation.

Indeed, the assumption of modularity* is often critical to understanding how feedback control enables robustness. This can be seen again in the example of the spider. In order to understand how the biomechanical relationships between the remaining legs need to change to preserve locomotion, it is necessary to first understand the isolated causal contribution of the destroyed leg. That leg will have played different roles in different motor behaviors—e.g. prey capture or web navigation. Characterizing those different roles—i.e. the isolated causal contributions of the leg—is necessary to understand how those motor behaviors can be maintained in the absence of the leg. For instance, whether it is a front, back, or middle leg that is missing will have consequences for the new motor patterns that need to be learned, and hence new causal relationships that need to be adopted, to maintain those behaviors.

These considerations are not limited to this example. Feedback control seems to be one of the primary mechanisms responsible for enabling functional robustness in a range of different systems. Of particular relevance, feedback control drives robustness in both neurons and genetic networks. Mitchell is, of course, aware of this. However, she comes to the conclusion that feedback is inconsistent with modularity. For instance, she writes, "[a]ny physical system with complex feedback mechanisms will be one in which we can expect modularity to fail. But we should not conclude that such systems don't involve true component causes" (Mitchell 2009, p.82). Unfortunately, she does not offer a definition of "true component causes" and indeed it is unclear what could be meant that isn't equivalent to modularity*. But as I've just argued, feedback and modularity* are not incompatible; to the contrary, modularity* is an important component of

adequate explanations of and investigations into systems that achieve robustness via feedback control.

This can be further illuminated by examining the other half of Swensen and Bean's (2005) study, which was briefly mentioned in chapter three. Recall that in their investigation of robustness of burst firing in Purkinje cells, Swensen and Bean (2005) performed two distinct interventions that targeted different timescales, and they found burst firing to be robust in both conditions. The first condition involved pharmacological blockade of sodium conductance with TTX, which is transient and occurs on very short timescales. The relevance of the short duration of the intervention is that it rules out second messenger processes occurring within the cell that may alter ion channel expression on the cell membrane. The second intervention involved a genetic knockout similar to those discussed in the previous section. In particular, the mice used were knockouts for the $Na_v1.6$ gene, which codes for the protein that constitutes a particular subtype of voltage-gated sodium channel.

Swensen and Bean conducted additional analyses to determine the mechanisms responsible for robustness in each condition, and they found evidence of distinct mechanisms that operate on two different timescales. The mechanism responsible for robustness under TTX intervention was described at length in the previous chapter. In the case of the genetic knockout study, which is more relevant to the issue at hand, Swensen and Bean found evidence of a completely different mechanism, compensating for the lack of sodium conductance by changing expression of other ion channels on the cell membrane.

They found that the main difference driving robustness of bursting in $Na_v1.6$ knockouts was an increase in calcium conductance. This is surprising for two reasons. First, recall that potassium conductance was the main current that changed in response to acute reductions in sodium

conductance in the TTX experiment. Potassium is positively charged and flows out of the cell, hyperpolarizing the membrane, whereas sodium is positively charged and flows into the cell, depolarizing the membrane. In the TTX experiment, Swensen and Bean found that as sodium conductance decreases, there is a compensating decrease in potassium conductance to maintain the small net influx of current necessary to drive burst firing. One might thus expect that in the knockout case, the changes in channel expression on the cell membrane would largely consist in a decrease in potassium channels proportional to the decrease in sodium channel expression. Yet this is not what Swensen and Bean found.

Second, recall also that one of the main drivers of increased potassium conductance in the TTX experiment was calcium-activated potassium current. Calcium influx can have, on net, either hyperpolarizing or depolarizing effects on membrane potential. While calcium entering the cell has a depolarizing effect, the hyperpolarizing response of calcium-activated potassium current is capable of overwhelming those depolarizing effects. However, Swensen and Bean found that this does not occur in the case of $Na_v1.6$ knockouts. This reflects plasticity in the coupling between calcium influx and calcium-activated potassium current.

In sum, there are two main changes that account for robustness of burst firing in the knockout case—increased calcium ion channel expression on the cell membrane and a decrease in the responsiveness of calcium-activated potassium current to the presence of calcium within the cell. Both these changes are likely the result of feedback control driving the cells toward burst firing in response to the decrease in sodium conductance due to the knockout.

This case reinforces two aspects of my arguments to this point. In the first place, note that the knocked-out sodium channels are modular* components of the system. This is all the more apparent due to the contrast between the two forms of intervention used by Swensen and Bean.

The transient, short-timescale intervention of TTX has a totally different effect on the system than the persistent, longer-timescale intervention of the gene knockout. What's more is that the modular* contribution of sodium conductance to action potential generation and burst generation was already well characterized prior to the study (Swensen and Bean 2003). This prior knowledge of the isolated causal contribution of sodium conductance played a significant role in guiding inquiry into the underlying mechanisms supporting robustness.

The second point reinforces points made in the previous section. Note that in the case of $Na_v1.6$ knockouts, if we zoom out to the level of the whole organism, the phenotypic change associated with the $Na_v1.6$ gene in the context of cerebellar function would likely be obscured. This is a case where (phenotypic) robustness occurs at an intermediate level within the system. If we were to follow Mitchell's prescriptions about causal investigation and explanation, we may well miss this effect. In other words, this case provides another instance where anomalous knockout results should not be taken at face value.

The preceding arguments show that functional robustness is consistent with modularity*. It does, nonetheless, carry significant consequences for accounts of causal explanation and causal inference. Different experimental interventions operate on different timescales, as we've just seen with genetic knockout and pharmacological intervention. These different techniques are capable of illuminating different causal mechanisms, as we've seen with the contrast between the two experiments from Swensen and Bean. The notion of ideal intervention in play in interventionist accounts of causation fails to capture this; ideal interventions are timescale insensitive. This seems to reflect a flaw in such accounts, or at least points to a set of issues where further work needs to be done.

Further consequences follow for methods involved in causal inference. Feedback control, by definition, entails cyclic causal structure—i.e. causal structure with bi-directional causal pathways linking at least two variables in the system. This is significant because it provides another reason to believe that systems that exhibit functional robustness may be unsuitable for analysis with the kinds of causal modeling techniques discussed in chapter three. Recall that that family of techniques rely on the assumption of acyclicity (hence, directed *acyclic* graphs—DAGs). While it may be informative to treat systems that exhibit robustness as acyclic on particular timescales in order to better understand particular modular* causes within the system, methods that assume acyclicity cannot capture the full dynamics of systems that involve feedback control.

4.6 Conclusion

In this chapter, I explored the consequences of functional robustness for theories of causation, explanation, and methods of causal investigation. I offered a tweak to the standard definition of modularity in interventionist theories of causation and argued that this amended notion of modularity* is preferable for a variety of reasons. I then considered Mitchell's argument that functional robustness undermines modularity*. I argued in contrast that there are multiple interpretations available for the anomalous genetic knockout that are the core of her argument. I showed that modularity*, as a methodological principle retains significant value in this domain by encouraging researchers to continue searching for new phenotypes associated with particular gene knockouts. I went on to argue that, in cases where functional robustness does in fact occur, it nonetheless is not incompatible with modularity*. I argued to the contrary that modularity* is a crucial aspect of causal investigations into the mechanisms that enable robustness. Finally, I

contended that the real consequences of functional robustness in these domains is its implication of causal cyclicity in the form of feedback control. I concluded by exploring some of the consequences of causal cyclicity for theories of causation and methods of causal investigation.

5.0 Dissertation Conclusion

In this dissertation, I have provided a novel account of multiple realization (MR) and explored its epistemic consequences. My aims in chapter 2.0 were largely positive. First, I provided an analysis of MR that moves away from positivist conceptions of explanation and reduction and operates instead within causal explanatory frameworks. Within such frameworks, I argued that MR can be construed as a thesis about the structure of causal explanations rather than a thesis about relations between kinds that figure into different taxonomic systems. The substance of the account of MR I developed is straightforward: multiple realization occurs when functions are stable—i.e. relevantly similar in their causal roles within some containing system—despite causally relevant differences in the ways the function is performed.

My second main aim in chapter 2.0 was to substantiate my account of MR through empirical examples of functional robustness in neuroscience. I argued that the traditional philosophical considerations that have surrounded MR (e.g. nomicity, projectibility, causal individuation) fail to adequately track important features of these empirical cases. This should perhaps be unsurprising given that traditional debates about MR are based on an outmoded framework of explanation and reduction in the special sciences.

One might worry, however, that interpolating MR into causal explanatory frameworks and aligning it with functional robustness might sacrifice much of what is philosophically interesting about MR in the first place. In particular, debates about MR have generally focused on its ability to secure the autonomy of psychology (or higher-level sciences, more generally) from neuroscience (or lower-level sciences, more generally). Within the positivist framework, this is the natural way to characterize the epistemic significance of multiple realization. In that framework,

multiple realization is cast as a thesis about the natural kind terms that figure into natural laws. If the natural kinds of a higher-level science are multiply realized by the kinds in the lower-level science, then the bridge principles that map between those kinds and are necessary for logical derivation are blocked. By contrast, in causal frameworks, rather than this comparatively thin epistemic thesis about autonomy, multiple realization instead implies a range of more nuanced epistemic consequences about causal discovery, the structure of causal explanation, how we proceed with causal investigation, and causal hypothesis testing. My aims in chapters 3.0 and 4.0 were to explore some of those epistemic consequences.

In chapter 3.0, I argued that systems that exhibit robustness will tend to violate causal faithfulness. I offered detailed analysis of the example of robustness of burst firing in Purkinje cells and showed how this example demonstrates a violation of faithfulness. I argued that the key feature driving robustness in this case is a form of parameter coupling that is well characterized in the causal dynamics of the system. This parameter coupling demonstrates how failures of the CFC can not only fail to be probability zero but can also be highly likely. I argued further that robustness is likely to be found in complex systems that maintain stable functions across timescales that exceed the lifespan of the component parts and processes that support those functions. I concluded by arguing that this likelihood of failures of faithfulness has significant consequences for both causal hypothesis testing and causal discovery.

In chapter 4.0, I explored the consequences of functional robustness for theories of causation, explanation, and methods of causal investigation. I offered a modification to the standard definition of modularity in interventionist theories of causation and argued that this amended notion of modularity is preferable for a variety of reasons. I then considered Mitchell's argument that functional robustness undermines modularity. I argued, in contrast, that there are multiple

interpretations available for the anomalous genetic knockout experiments that are the core of her argument. I showed that modularity, as a methodological principle, retains significant value in this domain by encouraging researchers to continue searching for new phenotypes associated with particular gene knockouts. I went on to argue that in cases where functional robustness does in fact occur, it nonetheless is not incompatible with modularity. I argued to the contrary that modularity is a crucial aspect of causal investigations into the mechanisms that enable robustness. Finally, I contended that the real consequence of functional robustness in these domains is its implication of causal cyclicity in the form of feedback control. I concluded by exploring some of the consequences of causal cyclicity for theories of causation and methods of causal investigation.

The epistemic consequences of robustness that I explored in chapters 3.0 and 4.0 diverge significantly from the traditional epistemic consequences that have been associated with MR. That should, however, be a welcome result. The shift in framing for MR that I advocated in chapter 1.0 influences not only where we should look for phenomena that exemplify MR but also what consequences MR should have for the epistemology of the sciences it factors into. To this end, I have only scratched the surface. There are a number of avenues of related and future research that are worth highlighting.

In particular, non-causal or non-mechanistic explanation has recently been a hot topic in philosophy of science (e.g. Batterman and Rice 2014, Chirimuuta 2018, Lange 2017, Ross 2015). Proponents of such explanations, generally, do not hark back to positivist views of explanation, but rather forge into interesting examples that test the limits of the causal frameworks of the new mechanists and interventionists that have been a focus of this dissertation. Other proponents of mechanistic explanation have also recently sought to extend the framework to network explanations (e.g. Bechtel 2017). The examples of functional robustness considered in this

dissertation are relevant to these debates. On one hand, robust functions seem to preclude causal analysis in terms of a single underlying causal mechanism. Indeed, the examples of robustness tend to show that there are multiple causal mechanisms that can support the same function (of course, the crux here is how mechanisms are individuated). Further, some of the same methods that proponents of non-causal explanation focus on, in particular dynamical models, have been used to understand what perturbations push robust systems out of their operable ranges. These models clearly play a role in homing in on the causal dynamics of these systems, but it is an interesting question whether the models should be regarded as explanatory. Much of this ground has already been covered in recent literature, but it is worth considering whether the examples of functional robustness bear on these debates in new and interesting ways.

Another topic that merits further exploration is that of causal inference in fMRI. I intimated in chapter 3.0 that failures of causal faithfulness are relevant to causal discovery in fMRI. To the extent that neural systems are robust and thus likely to generate failures of the CFC, this will confound a number of causal discovery algorithms that have currently been employed in efforts to recover functional connectivity from fMRI data.[45] It is often unclear exactly the extent to which these discovery algorithms assume the CFC. In some cases, however, it is transparent. For instance, the PC-algorithm begins with all possible edges connecting nodes in a graph, and then eliminates connections for variables that are independent or conditionally independent (thus transparently assuming the CFC).[46] Cashing out the full implications of failures of the CFC for such algorithms

[45] See Henry and Gates (2017) for an excellent review of these causal discovery algorithms and the extent of their application to fMRI data.
[46] See Spirtes and Glymour (1991) for details of the algorithm. See Joshi et al. (2010) for an application of the algorithm to fMRI data.

as well as their application to fMRI data is a potentially valuable extension of the research performed in chapter 3.0.

Finally, the issues raised toward the end of chapter 4.0 with respect to timescales and interventionism deserve further exploration. There are two issues here worth highlighting. The first can be tied in with the above considerations regarding the significance of robustness for fMRI research, though it also has more general implications. Specifically, there is a methodological concern that can be raised regarding temporal resolution and sampling rates of data used for causal modeling. Note, for instance, that the temporal resolution of fMRI tends to be on the order of several seconds; this is obviously a problem if the phenomena one is interested in involves feedback control processes that occur over shorter timescales (and in the case of neural systems it is entirely possible that they will).

The second point involving issues of timescales is that different experimental interventions often operate on different timescales. This is evident, for instance, in the difference between pharmacological knockout and genetic knockout. The notion of ideal intervention in play in interventionist accounts of causation fails to capture this because ideal interventions are framed in a way that is timescale insensitive. Ideal interventions are, of course, in-principle interventions, so one could argue that these differences in timescale of (in-practice) experimental techniques do not actually pose a challenge. However, these different interventions often reveal distinct causal mechanisms that occur on different timescales. This is again evident in the two distinct mechanisms sustaining Purkinje cell burst firing revealed in Swensen and Bean's (2005) study. In other words, the differences in timescale of intervention techniques does not simply reflect tradeoffs due to practical limitations of current technology, but instead actually tracks differences in the timescale on which phenomena in the world are stable. Interventionism is intended to

provide an account of the content of causal claims. To the extent that causal phenomena are stable on different timescales, some different forms of temporal indexing will be necessary to adequately characterize those phenomena. Building this kind of timescale sensitivity into the interventionist framework is a non-trivial task. For instance, it will involve not only characterizing the timescale over which the intervention occurs, but also the timescale over which other causal factors in the system are held fixed (or treated as constant). This is a bigger project than can even be thoroughly outlined here but exploring the implications of timescale relativity in causation may provide some important revisions to the interventionist framework.

Bibliography

Abbott, Larry F., and Sacha B. Nelson. 2000. “Synaptic Plasticity: Taming the Beast.” *Nature Neuroscience* 3 (11s): 1178.

Albensi, Benedict C. 2001. “Models of Brain Injury and Alterations in Synaptic Plasticity.” *Journal of Neuroscience Research* 65 (4): 279–83.

Albus, James S. 1971. “A Theory of Cerebellar Function.” *Mathematical Biosciences* 10 (1–2): 25–61.

Aldrich, John. 1989. “Autonomy.” *Oxford Economic Papers* 41 (1): 15–34. https://doi.org/10.1093/oxfordjournals.oep.a041889.

Andersen, Holly. 2013. “When to Expect Violations of Causal Faithfulness and Why It Matters.” *Philosophy of Science* 80 (5): 672–83.

Bach-Y-Rita, Paul. 2003. *Theoretical Basis for Brain Plasticity after a TBI. Brain Injury*. Vol. 17. https://doi.org/10.1080/0269905031000107133.

Ball, John M., Clarence C. Franklin, Anne-Elise Tobin, David J. Schulz, and Satish S. Nair. 2010. “Coregulation of Ion Channel Conductances Preserves Output in a Computational Model of a Crustacean Cardiac Motor Neuron.” *Journal of Neuroscience* 30 (25): 8637–49.

Barbaric, Ivana, Gaynor Miller, and T. Neil Dear. 2007. “Appearances Can Be Deceiving: Phenotypes of Knockout Mice.” *Briefings in Functional Genomics and Proteomics*. https://doi.org/10.1093/bfgp/elm008.

Batterman, Robert W., and Collin C. Rice. 2014. “Minimal Model Explanations.” *Philosophy of Science* 81 (3): 349–76. https://doi.org/10.1086/676677.

Bechtel, William. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

———. 2019. “Analysing Network Models to Make Discoveries about Biological Mechanisms.” *The British Journal for the Philosophy of Science* 70 (2): 459–84. https://doi.org/10.1093/bjps/axx051.

Bechtel, William, and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Bergsma, Wicher P., and Tamás Rudas. 2002. “Variation Independent Parameterizations of Multivariate Categorical Distributions.” In *Distributions With Given Marginals and Statistical Modelling*, edited by C.M. Cuadras, J. Fortiana, and J.A. Rodriguez-Lallena, 21–27. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-0061-0_3.

Bogen, Jim. 2004. “Analysing Causality: The Opposite of Counterfactual Is Factual.” *International Studies in the Philosophy of Science* 18 (1): 3–26. https://doi.org/10.1080/02698590412331289233.

Bullmore, Ed, and Olaf Sporns. 2012. “The Economy of Brain Network Organization.” *Nature Reviews Neuroscience* 13 (5): 336.

Cartwright, Nancy. 1999a. “Causal Diversity and The Markov Condition.” *Synthese* 121 (1): 3–27. https://doi.org/10.1023/A:1005225629681.

———. 1999b. *The Dappled World: A Study of The Boundaries of Science*. Cambridge: Cambridge University Press.

———. 2001. “Modularity: It Can - and Generally Does, Fail.”

———. 2002. “Against Modularity, the Causal Markov Condition, and Any Link between the Two: Comments on Hausman and Woodward.” *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/53.3.411.

———. 2004. “Causation: One Word, Many Things.” In *Philosophy of Science*, 71:805–19. https://doi.org/10.1086/426771.

———. 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Cconomics*. Cambridge: Cambridge University Press.

Chen, W, M A Kelly, X Opitz-Araya, R E Thomas, M J Low, and R D Cone. 1997. “Exocrine Gland Dysfunction in MC5-R-Deficient Mice: Evidence for Coordinated Regulation of Exocrine Gland Function by Melanocortin Peptides.” *Cell* 91 (6): 789–98. https://doi.org/10.1016/s0092-8674(00)80467-5.

Chirimuuta, M. 2018. “Explanation in Computational Neuroscience: Causal and Non-Causal.” *British Journal for the Philosophy of Science* 69 (3): 849–80. https://doi.org/10.1093/bjps/axw034.

Chowdhury, Rajiv, Sarah Stevens, Donal Gorman, An Pan, Samantha Warnakula, Susmita Chowdhury, Heather Ward, et al. 2012. “Association between Fish Consumption, Long Chain Omega 3 Fatty Acids, and Risk of Cerebrovascular Disease: Systematic Review and Meta-Analysis.” *BMJ* 345 (7881). https://doi.org/10.1136/bmj.e6698.

Cole, Donald C., Jill Kearney, Luz Helena Sanin, Alain Leblanc, and Jean Phillippe Weber. 2004. “Blood Mercury Levels among Ontario Anglers and Sport-Fish Eaters.” In *Environmental Research*, 95:305–14. https://doi.org/10.1016/j.envres.2003.08.012.

Craver, Carl F. 2001. “Role Functions, Mechanisms, and Hierarchy.” *Philosophy of Science* 68 (1): 53–74.

———. 2007. *Explaining the Brain. Explaining the Brain*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199299317.001.0001.

Cummins, Robert. 1975. “Functional Analysis.” *Journal of Philosophy* 72 (November): 741–64.

———. 1983. *The Nature of Psychological Explanation*. Cambridge, MA: MIT press.

———. 2000. “How Does It Work?" Versus" What Are the Laws?": Two Conceptions of Psychological Explanation.” In *Explanation and Cognition*, edited by Frank C. Kell and Robert Andrew Wilson, 117–44. Cambridge, MA: MIT press.

Fischer, Rico, and Franziska Plessow. 2015. “Efficient Multitasking: Parallel versus Serial Processing of Multiple Tasks.” *Frontiers in Psychology* 6 (1366). https://doi.org/10.3389/fpsyg.2015.01366.

Fodor, Jerry. 1997. “Special Sciences: Still Autonomous after All These Years.” *Noûs* 31 (June): 149–63. https://doi.org/10.1111/0029-4624.31.s11.7.

Glymour, Clark N., Peter Spirtes, and Kevin Kelly. 1987. *Discovering Causal Structure*. Orlando: Academic Press.

Goaillard, Jean-Marc, Adam L. Taylor, David J. Schulz, and Eve Marder. 2009. “Functional Consequences of Animal-to-Animal Variation in Circuit Parameters.” *Nature Neuroscience* 12 (11): 1424. https://doi.org/10.1038/nn.2404.

Goldman, Mark S., Jorge Golowasch, Eve Marder, and L.F. F Abbott. 2001. “Global Structure, Robustness, and Modulation of Neuronal Models.” *Journal of Neuroscience* 21 (14): 5229–38. https://doi.org/21/14/5229 [pii].

Golowasch, Jorge, Mark S. Goldman, L. F. Abbott, and Eve Marder. 2002. “Failure of Averaging in the Construction of a Conductance-Based Neuron Model.” *Journal of Neurophysiology* 87 (2): 1129–31.

Grashow, Rachel, Ted Brookings, and Eve Marder. 2010. “Compensation for Variable Intrinsic Neuronal Excitability by Circuit-Synaptic Interactions.” *Journal of Neuroscience* 30 (27): 9145–56.

Guo, Weinong, W. Edward Jung, Céline Marionneau, Franck Aimond, Haodong Xu, Kathryn A. Yamada, Thomas L. Schwarz, Sophie Demolombe, and Jeanne M. Nerbonne. 2005. “Targeted Deletion of Kv4.2 Eliminates Ito,f and Results in Electrical and Molecular Remodeling, with No Evidence of Ventricular Hypertrophy or Myocardial Dysfunction.” *Circulation Research* 97 (12): 1342–50. https://doi.org/10.1161/01.RES.0000196559.63223.aa.

Guo, Yi, Sushrut Jangi, and Michael A. Welte. 2005. “Organelle-Specific Control of Intracellular Transport: Distinctly Targeted Isoforms of the Regulator Klar.” *Molecular Biology of the Cell* 16 (3): 1406–16.

Haenisch, Britta, Stefan Herms, Manuel Mattheisen, Michael Steffens, Rene Breuer, Jana Strohmaier, Franziska Degenhardt, et al. 2013. “Genome-Wide Association Data Provide Further Support for an Association between 5-HTTLPR and Major Depressive Disorder.”

Journal of Affective Disorders 146 (3): 438–40. https://doi.org/10.1016/j.jad.2012.08.001.

Hamilton, J. P., M. Siemer, and I. H. Gotlib. 2008. "Amygdala Volume in Major Depressive Disorder: A Meta-Analysis of Magnetic Resonance Imaging Studies." *Molecular Psychiatry* 13 (11): 993–1000. https://doi.org/10.1038/mp.2008.57.

Hanwell, David, Toru Ishikawa, Reza Saleki, and Daniela Rotin. 2002. "Trafficking and Cell Surface Stability of the Epithelial Na+ Channel Expressed in Epithelial Madin-Darby Canine Kidney Cells." *Journal of Biological Chemistry* 277 (12): 9772–79.

Hausman, Daniel M., and James Woodward. 1999. "Independence, Invariance and the Causal Markov Condition." *British Journal for the Philosophy of Science* 50 (4): 521–83. https://doi.org/10.1093/bjps/50.4.521.

———. 2004. "Modularity and the Causal Markov Condition: A Restatement." *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/55.1.147.

Hempel, Carl G., and Paul Oppenheim. 1948. "Studies in the Logic of Explanation." *Philosophy of Science* 15 (2): 135–75.

Henry, Teague, and Kathleen Gates. 2017. "Causal Search Procedures for FMRI: Review and Suggestions." *Behaviormetrika* 44 (1): 193–225. https://doi.org/10.1007/s41237-016-0010-8.

Hock, C., G. Drasch, S. Golombowski, F. Müller-Spahn, B. Willershausen-Zönnchen, P. Schwarz, U. Hock, J. H. Growdon, and R. M. Nitsch. 1998. "Increased Blood Mercury Levels in Patients with Alzheimer's Disease." *Journal of Neural Transmission* 105 (1): 59–68. https://doi.org/10.1007/s007020050038.

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

Ito, Masao. 1989. "Long-Term Depression." *Annual Review of Neuroscience* 12 (1): 85–102.

Ito, Masao, and Masanobu Kano. 1982. "Long-Lasting Depression of Parallel Fiber-Purkinje Cell Transmission Induced by Conjunctive Stimulation of Parallel Fibers and Climbing Fibers in the Cerebellar Cortex." *Neuroscience Letters* 33 (3): 253–58.

Ito, Masao, Masaki Sakurai, and Pavich Tongroach. 1982. "Climbing Fibre Induced Depression of Both Mossy Fibre Responsiveness and Glutamate Sensitivity of Cerebellar Purkinje Cells." *The Journal of Physiology* 324 (1): 113–34.

Joshi, Anand A., Shantanu H. Joshi, Richard M. Leahy, David W. Shattuck, Ivo Dinov, and Arthur W. Toga. 2010. "Bayesian Approach for Network Modeling of Brain Structural Features." In *Medical Imaging 2010: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 7626:762607. SPIE. https://doi.org/10.1117/12.844548.

Kim, Jaegwon. 1992. "Multiple Realization and The Metaphysics of Reduction." *Philosophy and Phenomenological Research* 52 (1): 1–26.

———. 1999. “Making Sense of Emergence.” *Philosophical Studies* 95 (1): 3–36.

Kolb, Bryan, and Robbin Gibb. 1999. “Neuroplasticity and Recovery of Function Following Brain Injury.” *Cognitive Neurorehabilitation*, 9–25.

———. 2008. *Principles of Neuroplasticity and Behavior*. New York: Cambridge University Press.

Kuorikoski, Jaakko. 2012. “Mechanisms, Modularity and Constitutive Explanation.” *Erkenntnis* 77 (3): 361–80. https://doi.org/10.1007/s10670-012-9389-0.

Lange, Marc. 2017. *Because Without Cause*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190269487.001.0001.

Lim, Giselle P., Frédéric Calon, Takashi Morihara, Fusheng Yang, Bruce Teter, Oliver Ubeda, Norman Salem, Sally A. Frautschy, and Greg M. Cole. 2005. “A Diet Enriched with the Omega-3 Fatty Acid Docosahexaenoic Acid Reduces Amyloid Burden in an Aged Alzheimer Mouse Model.” *Journal of Neuroscience* 25 (12): 3032–40. https://doi.org/10.1523/JNEUROSCI.4225-04.2005.

Lindsey, James K. 1996. *Parametric Statistical Inference*. Oxford: Clarendon Press.

Machamer, Peter, Lindley Darden, and Carl F. Craver. 2000. “Thinking About Mechanisms.” *Philosophy of Science* 67 (1): 1–25.

MacLean, Jason N., Ying Zhang, Bruce R. Johnson, and Ronald M. Harris-Warrick. 2003. “Activity-Independent Homeostasis in Rhythmically Active Neurons.” *Neuron* 37 (1): 109–20.

Marder, Eve. 2011. “Variability, Compensation, and Modulation in Neurons and Circuits.” *Proceedings of the National Academy of Sciences* 108 (Supplement 3): 15542–48. https://doi.org/10.1073/pnas.1010674108.

Marder, Eve, and Jean-Marc Goaillard. 2006. “Variability, Compensation and Homeostasis in Neuron and Network Function.” *Nature Reviews Neuroscience* 7 (7): 563–74. https://doi.org/10.1038/nrn1949.

Marr, David. 1969. “A Theory of Cerebellar Cortex.” *Journal of Physiology* 202 (2): 437–70.

Meek, Christopher. 1995. “Strong Completeness and Faithfulness in Bayesian Networks.” In *Uncertainty in Artificial Intelligence: Proceedings of the 11th Conference*, edited by P. Besnard, 411–18. San Francisco: Morgan Kaufman Publishers.

Mitchell, Sandra D. 2008. “Exporting Causal Knowledge in Evolutionary and Developmental Biology.” In *Philosophy of Science*, 75:697–706. https://doi.org/10.1086/594515.

———. 2009. *Unsimple Truths : Science, Complexity, and Policy*. University of Chicago Press.

Mutter, Joachim, J. Naumann, R. Schneider, and H. Walach. 2007. “Mercury and Alzheimer’s Disease.” *Fortschritte Der Neurologie-Psychiatrie* 75 (9): 528–38. https://doi.org/10.1055/s-2007-959237.

Nerbonne, Jeanne M., Benjamin R. Gerber, Aaron Norris, and Andreas Burkhalter. 2008. “Electrical Remodelling Maintains Firing Properties in Cortical Pyramidal Neurons Lacking KCND2-encoded A-type K+ Currents.” *The Journal of Physiology* 586 (6): 1565–79.

Noppeney, Uta, Karl J. Friston, and Cathy J. Price. 2004. “Degenerate Neuronal Systems Sustaining Cognitive Functions.” *Journal of Anatomy* 205 (6): 433–42.

Pearl, Judea. 1998. “TETRAD and SEM.” *Multivariate Behavioral Research* 33 (1): 119–28.

———. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. https://doi.org/10.2307/3182612.

Pezawas, Lukas, Andreas Meyer-Lindenberg, Emily M. Drabant, Beth A. Verchinski, Karen E. Munoz, Bhaskar S. Kolachana, Michael F. Egan, Venkata S. Mattay, Ahmad R. Hariri, and Daniel R. Weinberger. 2005. “5-HTTLPR Polymorphism Impacts Human Cingulate-Amygdala Interactions: A Genetic Susceptibility Mechanism for Depression.” *Nature Neuroscience* 8 (6): 828–34. https://doi.org/10.1038/nn1463.

Piccinini, Gualtiero, and Carl Craver. 2011. “Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches.” *Synthese* 183 (3): 283–311.

Polger, Thomas W., and Lawrence A. Shapiro. 2016. *The Multiple Realization Book*. Oxford: Oxford University Press.

Prinz, Astrid A., Dirk Bucher, and Eve Marder. 2004. “Similar Network Activity from Disparate Circuit Parameters.” *Nature Neuroscience* 7 (12): 1345–52. https://doi.org/10.1038/nn1352.

Prinz, Astrid A, Vatsala Thirumalai, and Eve Marder. 2003. “The Functional Consequences of Changes in the Strength and Duration of Synaptic Inputs to Oscillatory Neurons.” *Journal of Neuroscience* 23 (3): 943–54.

Putnam, Hilary. 1975. “Philosophy and Our Mental Life.” In *Mind, Language and Reality: Philosophical Papers, Volume 2*, 291–303. Cambridge: Cambridge University Press.

Ransdell, Joseph L., Satish S. Nair, and David J. Schulz. 2012. “Rapid Homeostatic Plasticity of Intrinsic Excitability in a Central Pattern Generator Network Stabilizes Functional Neural Network Output.” *Journal of Neuroscience* 32 (28): 9649–58.

Ransdell, Joseph L, Satish S Nair, and David J Schulz. 2013. “Neurons within the Same Network Independently Achieve Conserved Output by Differentially Balancing Variable Conductance Magnitudes.” *Journal of Neuroscience* 33 (24): 9950–56. https://doi.org/10.1523/JNEUROSCI.1095-13.2013.

Reichenbach, Hans. 1956. *The Direction of Time*. Los Angeles: University of California Press.

Ross, Lauren N. 2015. “Dynamical Models and Explanation in Neuroscience.” *Philosophy of Science* 82 (1): 32–54. https://doi.org/10.1086/679038.

Rossi, Peter H., Richard A. Berk, and Kenneth J. Lenihan. 1980. “Money, Work and Crime: Some Experimental Results.” New York: Academic Press.

Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Scheines, Richard, Peter Spirtes, Clark N. Glymour, Christopher Meek, and Thomas Richardson. 1998. “Reply to Comments.” *Multivariate Behavioral Research* 33 (1): 165–80. https://doi.org/10.1207/s15327906mbr3301_8.

Scherk, Harald, Oliver Gruber, Patrick Menzel, Thomas Schneider-Axmann, Claudia Kemmer, Juliana Usher, Wolfgang Reith, Jobst Meyer, and Peter Falkai. 2009. “5-HTTLPR Genotype Influences Amygdala Volume.” *European Archives of Psychiatry and Clinical Neuroscience* 259 (4): 212–17. https://doi.org/10.1007/s00406-008-0853-4.

Schofield, Paul N., Robert Hoehndorf, and Georgios V. Gkoutos. 2012. “Mouse Genetic and Phenotypic Resources for Human Genetics.” *Human Mutation*. https://doi.org/10.1002/humu.22077.

Schulz, David J, Jean-Marc Goaillard, and Eve Marder. 2006. “Variable Channel Expression in Identified Single and Electrically Coupled Neurons in Different Animals.” *Nature Neuroscience* 9 (3): 356.

Schupbach, Jonah N. 2018. “Robustness Analysis as Explanatory Reasoning.” *The British Journal for the Philosophy of Science* 69 (1): 275–300.

Shapiro, Lawrence A. 2000. “Multiple Realizations.” *Journal of Philosophy* 97 (12): 635–54.

———. 2004. *The Mind Incarnate*. Cambridge, MA: MIT press.

Shapiro, Lawrence A., and Thomas W. Polger. 2012. “Identity, Variability, and Multiple Realization in the Special Sciences.” In *New Perspectives on Type Identity the Mental and the Physical*, edited by Simone Gozzano and Christopher S. Hill, 264–87. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511687068.014.

Spirtes, Peter, Clark N. Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. *SpringVerlag Lectures in Statistics*. Cambridge, MA: MIT press.

Spirtes, Peter, and Clark N Glymour. 1991. “An Algorithm for Fast Recovery of Sparse Causal Graphs.” *Soc Sci Comput Rev* 9 (1): 62–72. http://repository.cmu.edu/philosophy.

Spirtes, Peter, Richard Scheines, Clark N. Glymour, Thomas Richardson, and Christopher Meek. 2004. “Causal Inference.” In *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, edited by David Kaplan, 448–77. SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.

Steel, Daniel. 2006. “Homogeneity, Selection, and the Faithfulness Condition.” *Mind Mach* 16: 303–17.

———. 2008. “Comment on Hausman & Woodward on the Causal Markov Condition.” *The British Journal for the Philosophy of Science*. Oxford University PressThe British Society for the Philosophy of Science. https://doi.org/10.2307/3541659.

———. 2010. “Cartwright on Causality: Methods, Metaphysics and Modularity - Hunting Causes and Using Them: Approaches in Philosophy and Economics, Nancy Cartwright. Cambridge University Press, 2008, x + 270 Pages.” *Economics and Philosophy* 26 (1): 77–86. https://doi.org/10.1017/s0266267110000064.

Stjepanović, D., V. Lorenzetti, M. Yücel, Z. Hawi, and M. A. Bellgrove. 2013. “Human Amygdala Volume Is Predicted by Common DNA Variation in the Stathmin and Serotonin Transporter Genes.” *Translational Psychiatry* 3. https://doi.org/10.1038/tp.2013.41.

Strata, Piergiorgio. 2009. “David Marr’s Theory of Cerebellar Learning: 40 Years Later.” *The Journal of Physiology* 587 (Pt 23): 5519.

Stromatias, Evangelos, Daniel Neil, Michael Pfeiffer, Francesco Galluppi, Steve B. Furber, and Shih-Chii Liu. 2015. “Robustness of Spiking Deep Belief Networks to Noise and Reduced Bit Precision of Neuro-Inspired Hardware Platforms.” *Frontiers in Neuroscience* 9 (222): 1–14.

Swensen, Andrew M., and Bruce P. Bean. 2003. “Ionic Mechanisms of Burst Firing in Dissociated Purkinje Neurons.” *Journal of Neuroscience* 23 (29): 9650–63.

———. 2005. “Robustness of Burst Firing in Dissociated Purkinje Neurons with Acute or Long-Term Reductions in Sodium Conductance.” *Journal of Neuroscience* 25 (14): 3509–20. https://doi.org/10.1523/JNEUROSCI.3929-04.2005.

Taylor, Adam L., Jean-Marc Goaillard, and Eve Marder. 2009. “How Multiple Conductances Determine Electrophysiological Properties in a Multicompartment Model.” *Journal of Neuroscience* 29 (17): 5573–86.

Williams, T A, and L M Nagy. 2001. “Developmental Modularity and the Evolutionary Diversification of Arthropod Limbs.” *The Journal of Experimental Zoology* 291 (3): 241–57. https://doi.org/10.1002/jez.1101.

Woodward, James. 1998. “Causal Independence and Faithfulness.” *Multivariate Behavioral Research* 33 (1): 129–48.

Woodward, James F. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

———. 2008. “Invariance, Modularity, and All That.” In *Nancy Cartwright’s Philosophy of Science*, edited by & L. Bovens S. Hartman, C. Hoefer, 198–237. New York: Taylor & Francis.

Wrinn, K. M., and G. W. Uetz. 2007. “Impacts of Leg Loss and Regeneration on Body Condition, Growth, and Development Time in the Wolf Spider Schizocosa Ocreata.” *Canadian Journal of Zoology* 85 (7): 823–31. https://doi.org/10.1139/Z07-063.

Yamazaki, Tadashi, and William Lennon. 2019. “Revisiting a Theory of Cerebellar Cortex.” *Neuroscience Research*.

Zeisel, Hans. 1982. “Disagreement over the Evaluation of a Controlled Experiment.” *American Journal of Sociology* 88 (2): 378–89.

Zhang, Jiji, and Peter Spirtes. 2008. “Detection of Unfaithfulness and Robust Causal Inference.” *Minds and Machines* 18 (2): 239–71.