January 2011

# The Manipulated Mechanism: Towards an Account of the Experimental Discovery of Mechanistic Explanations

Donald Goodman-Wilson
*Washington University in St. Louis*

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Philosophy

Dissertation Examination Committee:
Carl Craver, Chair
Dennis Des Chenes
Frederick Eberhardt
Gillian Russell
Thomas Sattig
Lawrence Snyder
Kurt Thoroughman

# THE MANIPULATED MECHANISM

# TOWARDS AN ACCOUNT OF THE EXPERIMENTAL DISCOVERY OF MECHANISTIC EXPLANATIONS

by

Donald Edward Goodman-Wilson

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2011

Saint Louis, Missouri

**Abstract**

Recent work in the philosophy of biology has sought after an account of mechanistic explanation. Biologists frequently encounter causal relationships that beg for explanation. For example, genes appear to encode for particular phenotypes. How does gene expression *work*? Biologists posit *mechanisms* to explain the link between cause and effect. Thus, gene expression would be explained by an appeal to a complex mechanism linking the gene to the phenotype, as such an appeal will provide answers to broad ranges of "how" and "why" questions about the causal relationship, and predict novel effects.

Here, I focus on a recent problem raised for mechanistic explanation. Mechanism discovery is an inferential process which takes empirical data as premises, and produces a causal model of a mechanism as the conclusion. Such an inferential process requires rules, yet few accounts of mechanistic explanation attempt to provide them. Such inferential rules could be used to answer related normative questions facing accounts of mechanistic explanation. In particular, they can be brought to bear on questions of explanatory relevance: Which components are part of the mechanism, and how can we know? and questions of explanatory adequacy: When is a mechanistic explanation a *good* explanation? I argue that a formal account of mechanistic explanation grounded in a manipulationist account of causation can answer these kinds of question.

A thoroughgoing defense of my account, however, requires that I defend its assumptions. Among the assumptions is the highly contentious principle known as 'modularity'. Modularity is the claim that we must be able to independently manipulate each of the various components in a mechanism. The final chapters of my dissertation focus on a thoroughgoing defense of modularity against claims that it is frequently violated, conceptually intractable, or simply inapplicable to especially biological systems.

**Acknowledgements**

It is important that I acknowledge those who have supported me and contributed to this work. Without them, this dissertation would not exist. The following have my heart-felt thanks.

Carl Craver, my dissertation director, for being a sounding board for my ideas since I arrived at Washington University, for reading and commenting on paper after paper, draft after draft, on the same topic for six years straight, and for his guidance and advice generally.

Frederick Eberhardt, for keeping me honest, and in particular for helping me salvage a chapter or two from what was originally my first qualifying paper.

Dennis Des Chene, for his advice on writing (especially introductions), and for his encouragement in things academic and otherwise.

Gillian Russell, Thomas Sattig, Lawrence Snyder, and Kurt Thoroughman—the remainder of my committee—, for their time and thoughtful questions and comments at my defense.

The Department of Philosophy at Washington University in St Louis, for funding me through to the end, and especially Eric Brown and Kit Wellman who argued the case for my continued funding to the dean.

Lynn Holt of Mississippi State University, for suggesting that philosophy was what I *really* wanted to do.

Sarah Robbins and David Bauman, for standing by me in dark times and light.

Amy Goodman-Wilson, for everything.

To Acadia, with all that follows.

# Contents

# List of Tables

# List of Figures

# Toward a Normatively and Descriptively Adequate View of Mechanisms


## Abstract

In 1948, Hodgkin and Huxley demonstrated a curious relationship between the voltage and current inside a neuron just before it fires (the so-called 'action-potential'). They observed that, when one applies a certain fixed voltage across the cell membrane of a prepared squid giant axon, the current across the membrane changes in accordance with Ohm's Law[1]—but only briefly. Then, the voltage across the membrane drops precipitously (whatever the applied voltage might be), and current runs first out, then into the cell before the membrane voltage stabilizes at a steady value. This curious relationship is not observed in ordinary metallic conductors, whose behavior is described entirely by Ohm's Law. Something far more complex was going on inside neurons.

Although they refused to speculate about the mechanism responsible for this curious relationship in their report (Hodgkin, Huxley, & Katz, 1952; Hodgkin & Huxley,


[1] $R = V/I$; for a fixed resistance, voltage and current are inversely proportional.

1952a,b,c,d), they did provide a mathematical model of the relationship, a fourth-degree polynomial approximation over the data. Yet, the success of the experiments necessary to collect this data required some understanding of the mechanism that they were studying. Indeed, they clearly saw that this non-linear response was generated by a pair of feedback loops, one positive (driving the voltage far down), and one negative (returning the system to a steady state). They could see this because the feedback loops had to be broken for accurate measurements to be taken in the first place. They used a device called a voltage clamp, which (as the name implies) permits the experimenter to hold the membrane voltage at a fixed value, preventing it from shooting downwards, by breaking the feedback loops.

How could they know how to break the feedback loops, without at least some understanding of the mechanism for this curious relationship? Indeed, they privately speculated at length about possible mechanisms for the action potential (Huxley, 2002), and although none of them quite predicted the actual data recorded in their experiments these speculations did crucially inform their calculations of the mathematical model they did publish. None of the speculations were ever published, and Hodgkin and Huxley were careful to point out that the model they developed should not be viewed as an explanation of the data: It did not in their view, embody a mechanism. Why not? Craver (2008) points out that Hodgkin and Huxley do provide additional evidence (an electrical model, evidence that the current comprised the movement of sodium and potassium ions, *&c.*) that could be used to give their mathematical model a causal interpretation, and hence to take a significant step forward in giving a mechanistic explanation. And yet, Hodgkin and Huxley did *not* take advantage of this evidence, steadfastly insisting that the model they developed could not itself stand as an explanation. Indeed, nearly twenty years would pass before researchers were willing to give the Hodgkin and Huxley model a causal interpretation

Hille, Armstrong, & MacKinnon (1999).

This reluctance is puzzling. Given the causal knowledge they must have had, why did Hodgkin and Huxley refrain from making claims about the mechanism, or even restricted claims about the causal structure of the mechanism, in their report? Although they were rightly concerned that they had inadequate evidence that any of the variables in their model referred to real properties of real things in the neuron cell, they certainly had experimental evidence that the mathematical models could, nevertheless, adequately describe not just correlations among voltage and current, but causal connections. Yet they did not take this step. What additional evidence did they and later researchers believe they required to make the leap from a functional description to mechanical explanation?

## Selecting an Explanatory Framework

Present models of explanation, it seems to me, cannot account for this reluctance.[2] Hempel & Oppenheim's (1948) D-N model of explanation (as Craver (2008) points out) entails that the Hodgkin and Huxley model *is* in fact a good explanation. The D-N model says that an explanation is a kind of sound argument with natural laws as premises and the explanandum as the conclusion. And, indeed, the Hodgkin and Huxley model can be read as a set of laws (or, more broadly, lawlike generalizations) from which we can derive the complex relationship between voltage and current to be explained. But, whatever other problems the D-N model has in this case, this conclusion is at variance with Hodgkin and Huxley's own assessment.

More recent causal accounts of explanation cannot account for their reluctance either. The conserved-quantity or physical account of causation (Salmon (1997); Dowe

[2]What follows is not meant to convince the reader that I am right so much as to simply introduce my own position, and offer some evidence that it is *prima facie* tenable.

(2000) claims that giving a causal explanation involves tracing the transmission of a conserved quantity (*e.g.* energy or mass) in the etiology for the explanandum phenomenon. And yet, insofar as the Hodgkin and Huxley model *does* describe the transmission of a conserved quantity, namely charge, it qualifies as a causal explanation. Again, this is at variance with the actual position of the model's authors.

A counterfactual account of causal explanation gets us a little closer. On this view, an explanation consists in true counterfactuals that describe causal relations. The Hodgkin and Huxley model does in fact describe a range of true causal counterfactual statements, but it also supports a much broader range of non-causal counterfactuals[3]. Thus, on the counterfactual view, additional evidence is needed to constrain the range of counterfactual claims supported by the model. Indeed, Hodgkin and Huxley agree, when they claim that their model picks out a broad class of possible mechanisms. But Hodgkin and Huxley also worry that many of the terms in the model do not refer to real components, a worry not accounted for by the counterfactual account.

Recently, philosophers have begun to analyze mechanisms as a special kind of causal explanation, and to understand the role that mechanisms play in the practice of biology (*e.g.*, Bechtel & Richardson, 1993; Machamer, Darden, & Craver, 2000; Glennan, 2002). Craver (2008) takes the Hodgkin and Huxley model, and Hodgkin and Huxley's reaction to it, as evidence for this view of explanation. Mechanistic explanation claims that explanation involves analyzing the phenomenon into distinct parts and showing how those parts interact to produce the phenomenon. On this view, the Hodgkin and Huxley model is not explanatory, because it is a formal representation: It does not pick out component parts of the mechanism. While the model does say something about mathematical relations that hold within the mechanism—and so counts only as what Craver calls a mechanism sketch—, it does not pick out

[3]Including what Lewis called 'backtracking counterfactuals', or a counterfactual whose evaluation requires holding the present fixed and evaluating how the past must have differed.

specifically causal interactions because it does not pick out the parts, the causal relata. More is needed for the model to count as an explanation: To be explanatory, the model must be given an interpretation such that the variables can be read as pick out parts, and the functional relations as picking out causal relations among those parts. So far, so good: This account does, to this point, a good job of accounting for Hodgkin and Huxley's understanding. And yet, although on the mechanistic view Hodgkin and Huxley had enough puzzle pieces to form a mechanism sketch, they chose explicitly to not take that step. The mechanistic account says that they would have been justified in doing so, but they did not believe that they were. Mechanistic explanation doesn't yet quite capture the norms of explanation in play in this historical episode.

## The Need for Rapprochement

As these mechanistic analyses mature, philosophers of biology have come to ask questions about mechanistic explanation that require a normative framework: What makes a good mechanistic explanation? How can we know when we have a correct mechanistic explanation? One way to answer these questions is to look towards causal modeling techniques (*e.g.*, Pearl, 2000; Spirtes, Glymour, & Scheines, 1993). These mathematical models provide a set of principles for inferring causal relations, and a set of norms for determining when these principles can be reasonably applied. Thus, causal modeling appears a good place to begin looking for answers. Yet, although there are descriptive accounts of mechanisms, and there are mathematical models of causal systems, neither alone is complete as an adequate account of mechanistic explanation, and there is no currently available account that does justice to the two of them.

Part of the reason that there is currently no account that bridges these two fields is that many philosophers of biology are shy of abstract, normative claims, driven to

avoidance by a dilemma: Normativity can be founded in either *a priori* principles, or in *a posteriori* principles. But a reliance on *a priori* principles tends to distort actual scientific practice (as evidenced by the failures of logical empiricism); and a reliance on *a posteriori* principles commits the is-ought fallacy. Thus, one might think, there can be no normative philosophy of biology.

But this dilemma is false. The worry that *a priori* principles distort practice is, I think, largely a reaction to the failures of logical empiricism: Burned once, philosophers of biology would rather not cross back into prescriptive territory. I think that we *can* construct an account of mechanistic explanation that uses *a priori* principles as a normative foundation, yet with minimal distortion of actual practice. The crucial step is to recognize the central lesson of the naturalistic turn: Don't just bring your normative account to science and see how it measures up, but use actual scientific practice to inform the selection of *a priori* principles (as there are many to choose from) that are most likely to avoid undue distortion of actual practice.[4] The way, in other words, of avoiding the descriptivist dilemma is to bring the descriptive work and the normative work into rapprochement.

What must such a rapprochement accomplish? On one hand, although the extant accounts of mechanistic explanation do a fine job of describing many of the explanations that biologists give, such are unable to evaluate these explanations. These descriptive accounts lack an account of explanatory relevance—a measure of how central a component or event is to a particular explanation—, and they lack any kind of apparatus for constructing and evaluating the experiments, and the inferences made from those experiments, that led to those very mechanistic explanations. Worth noting is that Bechtel & Richardson (1993) and Darden (2001); Craver & Darden (2001);

[4]Some distortion is inevitable, at least insofar as biologists do not deploy sound reasoning in their evaluation of mechanistic explanations. But, if there were no room for critique and improvement, there would be no need for philosophical accounts of explanation in the first place.

Darden (2002, 2006) do, of course, have an account of the strategies that biologists *do* use for discovering experiments, but that these accounts deliberately steer away from making the additional claim that these strategies are the *right* strategies, or from offering principles by which one could evaluate these strategies generally. On the other hand, satisfactory solutions to these problems cannot simply deny actual scientific practice. It is easy to apply purely abstract models of causal reasoning to scientific reasoning, without paying attention to the physical details that interest biologists. A complete account of scientific explanation should not rely on pure formalism, but should have a story as to why biologists do concern themselves with the particular details.

Recent work on the problem has focused on crafting a rapprochement founded on quantitative accounts of causal explanation, and in particular, on manipulationism. Manipulationism claims that one of the defining characteristics of a causal relation is that causes can be used to manipulate their effects. Thus, we can explain an effect by an appeal to how that effect would have been different, had its causes been different. Woodward (2002), for example, claims that a mechanism is a collection of entities bound by relationships of manipulability; a causal network, in essence. But such a definition is overly broad: Nearly any causal system can be described as a mechanism on this view. But not every causal system *is* a mechanism. The moon's position in the sky has an effect on my dog's gait (as it does on the tides), yet such a causal network is likely not the sort of thing of interest to mammalogists. By failing to attend to important qualitative details, Woodward's account can only fuel the worries of the descriptivists.

Craver (2007) has attempted a more naturalistic approach to building an account of mechanistic explanation from manipulationism. Craver distinguishes two modes of causal explanation: Constitutive and etiological. Where etiological explanations ex-

plain by appeal to the antecedent causes of a phenomenon, constitutive explanations explain by appeal to the causal interactions among the parts and components of the phenomenon. Most mechanistic explanations offered in biology and neuroscience, he observes, are of the constitutive kind. Thus, Craver identifies a particular subset of causal explanations as specifically mechanistic. Craver harnesses the manipulationism to give an account of constitutive relevance; We can know an entity is part of a mechanism if it is related to the mechanism as part to whole (that is, is spatially located within the physical bounds of the mechanism), and when we can manipulate the mechanism by intervening into the entity and *vice versa*.

But where manipulationism claims that manipulability is a hallmark of causal relations, Craver has now added that manipulability is *also* a hallmark of constitutive relevance. As a result, Craver is faced with a dilemma. On the one hand, perhaps constitutive relevance is a species of causal relation. But this view commits Craver to that mechanisms engage in inter-level causation with their components—if we can manipulate a component by manipulating the mechanism, it is because the mechanism is a cause of the component. But Craver explicitly—and I think rightly—denies this view. On the other hand, Craver risks creating epistemological confusion: Given a relationship of mutual manipulability, how are we to know whether we face a relationship of constitutive relevance, or a causal feedback loop? Although I leave off argument for Chapter 3, I do not think an additional appeal to spatial (part-whole) relations can provide the necessary distinction. If I am right, then, on this hand, Craver's view cannot quite accomplish what it sets out to.

As I will argue, the way out of this dilemma is to take a different approach to the formal description of mechanisms. Rather than thinking of mechanisms as constitutive, we should use etiological descriptions of mechanisms. By doing so, what were once constitutive relationships become causal relationships, and the dilemma

can be avoided by losing this distinction.

## Towards a Descriptively and Normatively Adequate Account of Mechanism

Manipulationism is a formal account that offers a set of principles and constraints on when we can interpret a graph or a set of equations as representing specifically causal relations. Mechanisms comprise a bounded set of components[5] and the interactions or activities that bind them together into a coherent, stable, productive whole. How should components and activities be represented by a graph or an equation? The accounts of Bechtel & Richardson (1993) or Machamer, Darden, & Craver (2000) have no account of relevance to offer, no principles for representing mechanisms. Indeed, it seems no author has yet taken seriously the matter of finding the right mapping between the elements of the descriptive accounts of mechanism with the formal elements of manipulationism. Woodward (2002) and Craver (2007) have the first tentative steps forward, but there is much work to do. This dissertation focuses on taking the next step forward.

Such a semantics for mechanisms would allow us to make sense of Hodgkin and Huxley's decision. Their model of the current-voltage relationship comprises variables with clear physical interpretations (variables that represent the movement of sodium ions, for example), and theoretical terms with no clear referent, inserted only to make the model fit the data. Their reluctance to view their model as explanatory seems to hinge on the lack of a clear mapping between these theoretical terms and any possible mechanism component. Without a clear mapping, they had no way to constrain the possible range of causal interpretations on the model, and hence no clear way to map their model onto any kind of mechanism, possible, plausible, or otherwise.

[5] Although how they are bounded varies widely from account to account.

Thus, only once we have an account of mechanism that includes normative principles for mapping formal elements of mathematical models with real, concrete elements in biological systems can we begin to make sense of the inferential decisions that scientists like Hodgkin and Huxley and their successors have made, and to judge such decisions and their outcomes as proper or improper.

The first goal of this dissertation is to complete a rapprochement between descriptive accounts of mechanism with formal account of causal inference, by proposing and defending such a mapping. I will argue that the various descriptive accounts of mechanism provide sufficient structural and interpretive constraints to ensure a unique mapping of manipulationist elements to mechanistic elements. Formal models comprise causal relations and causal relata. Mechanisms link cause to effect *as* cause to effect; mechanisms form hierarchies. I will argue that these two features of mechanisms rule out certain classes of causal models as not describing mechanisms. Mechanisms comprise entities, activities, and interactions. I will argue that this feature of mechanisms limits the possible range of interpretations of the relata and relations in mathematical models; namely, such that the causal relata are component-activity pairs, and the causal relations are relations of interactivity. Finally, I will argue that, for any adequate description of a mechanism, these constraints are sufficient to pick out exactly one manipulationist causal model and interpretation of that model; that there is a one-to-one relationship between mechanism descriptions and mechanism models. The end product is a set of principles for bridging properties of formal models with properties of real biological mechanisms, and with real experimental techniques, completing the rapprochement. I will argue, in closing, that the rapprochement presented in this dissertation is descriptively adequate by applying it to a range of historical cases of mechanism discovery, and showing how the principles of the rapprochement account for what would otherwise seem peculiar features of the experiments described.

The second goal of this dissertation is to defend my account against a recent and potentially devastating objection to manipulationism. Because the rapprochement is founded on manipulationism, and manipulationism is committed to an idea called *modularity*. Modularity is, roughly, the idea that it must be possible to intervene into any component of a mechanism independently of the rest, a feature called *modularity*. Yet, precious few mechanisms in biology exhibit this property. Therefore, since manipulationism requires modularity (Glymour, 2004), and the rapprochement require manipulationism, a devastating attack on modularity is a devastating attack on the rapprochement.

In defense of a manipulationist rapprochement, I will provide two arguments. I will first argue that certain of the characterizations of modularity offered by its critics are either uncharitable, and false: They are far stronger, at any rate, than manipulationism requires. I will second argue that the remaining objections to modularity hang on some part of the modularity principle that can be safely dropped, at least in the context of mechanistic explanation, and that dropping this part of modularity does away with the objection without undermining manipulationism more broadly.

## Plan of the Dissertation

This dissertation is divided into three broad parts: The first part contains an overview of causal modeling and mechanistic explanation. The second part motivates and presents my quantitative account of mechanism, which I call *the manipulated mechanism*. The third part defends the manipulated mechanism against the criticism that my account depends crucially on a contentious principle called *modularity*—a principle regularly violated by biological mechanisms.

**Chapter 1** I begin the rapprochement by considering in detail both elements that

I wish to bring together. In the first chapter, I consider the formal element, offering a motivation for the inferential tools that causal modeling can provide philosophers of biology. I give a brief précis of the machinery of causal models, and in particular Woodward's manipulationist framework for reasoning in causal models. Causal modeling techniques include the use of causal graphs and causally interpreted systems of equations. I present the basic tools for bridging these different kinds of models, and drawing causal inferences from them. I also introduce the manipulationist principles of invariant generalizations, interventions, and modularity. A normative account of mechanism that relies on causal modeling and manipulationism as a quantitative foundation will draw heavily upon these tools.

**Chapter 2** Having examined the quantitative element of the rapprochement, I turn in this chapter to consider the descriptive element of the rapprochement. This chapter presents three different descriptive accounts of mechanism, due to Bechtel & Richardson (1993), Glennan (1996, 2002), and Machamer, Darden, & Craver (2000). Mechanisms link cause to effect via the complex interactions of a number of component parts, and have well-defined start conditions (on the cause) and end conditions (on the effect). Mechanisms are bounded, and can be hierarchically decomposed into sub-mechanisms. These common characterizations of mechanism will be crucial in crafting my rapprochement in Chapter 4.

**Chapter 3** With both elements of a rapprochement in hand, I turn to examine how these elements have been deployed in earlier steps toward a rapprochement. Woodward (2002) and Craver (2007) have both offered independent extensions of mechanistic explanation that attempt to provide a quantitative basis for

representing and reasoning about mechanisms. Both authors have argued that a rapprochement of mechanism and manipulationism holds promise to solve both the external and the internal challenges to a complete account of mechanistic explanation; yet these two attempts are not without difficulties. Woodward fails to capture the descriptive strength of the qualitative accounts of mechanisms; Craver fails to capture the prescriptive strength of manipulationism.

**Chapter 4** But these failures do not mean rapprochement is impossible; indeed, they point the way to improvement. My diagnosis is that previous authors have not yet taken seriously the challenge of mapping the formal elements of causal models to the real elements of mechanisms. In this chapter, I develop a set of constraints that the descriptive accounts of mechanism place on the formal elements of manipulationism—a semantics for interpreting a causal model as a biological mechanism. I begin by observing that causal models are blank slates, open to interpretation. On the other hand, the qualitative accounts of literature are concrete and specific—indeed, this is their very descriptive strength. I canvass the qualitative accounts for particular features that contribute to that descriptive strength. Mechanisms link cause to effect; mechanisms are bounded; mechanisms link entities and activities. I treat these features as constraints on the possible interpretations of causal models; thus, only causal models that link a cause to an effect, that have a clearly articulated mechanism bounding principle, and that relate entities and activities can count as models of mechanism. I also observe that purely etiological accounts of mechanism can avoid Craver's struggle to fit a constitutive concept of mechanism with a flat modeling technique. The qualitative constraints, in addition to constraining the possible causal structure of a mechanism, also point the way to a purely etiological account of mechanistic relevance, and thus allows us to side-step the worries about

relevance that drove Craver to adopt a hierarchical model in the first place.

These qualitative constraints and an etiological mechanistic relevance principle, when conjoined to the manipulationist account of causation, constitute the manipulated mechanism, a complete quantitative account of mechanism. This account holds the descriptive strength of the qualitative accounts, and the prescriptive strength of manipulationism.

**Chapter 5** However, because the manipulated mechanism leans so heavily on manipulationism, it is subject to a peculiar but worrisome objection: Manipulationism, roughly, requires that causal structures comprise independent components that can be removed or altered independently of the other components (much like the components in a mass-produced automobile). In this chapter, I examine why manipulationism requires the modularity principle and I discriminate the principle at stake with several similar principles that share the name 'modularity'. Specifically, manipulationism requires that when we intervene into any component in a mechanism, that that intervention must be probabilistically independent (uncorrelated) with any component that is not an effect of our intervention—a concept Woodward has formalized in a principle he calls **(PM)**. But, the objections observe, very precious few mechanisms in biology (or elsewhere) are modular in this way. So the manipulated mechanism will fail to account for these common biological mechanisms. There are three variants on this objection, and I must address each to defend the manipulated mechanism.

**Chapter 6** The first variant on this objection, which I address in this chapter, ties modularity to the causal Markov condition, which states that, for each component, that component's causes screen it off (render it conditionally independent) from it's non-effects. But the causal Markov condition is frequently violated

by non-deterministic systems that exhibit all-or-none responses—many medical syndromes fit the bill—and by linear systems that depend on not just the immediately prior state of the system, but all prior states—such as population growth models. Such cases violate the causal Markov condition. But this objection fails, for two reasons. First, I argue that some mechanisms that violate the Markov condition do *not* violate **(PM)**. Second, I argue that **(PM)** embodies two distinct claims, one of which bears no relationship to modularity as discussed in Chapter 5; I argue that the remaining mechanisms that violate the Markov condition only violate this extraneous condition of **(PM)**. For this reason, I close the chapter with an argument that we should weaken our formulation of modularity to exclude this condition. What remains is a condition I call **(PMa)**.

**Chapter 7** The second variant of the objection, addressed in this chapter, understands modularity to require a mechanism to exhibit a feature I call *modular independence*, that each component of the mechanism must present some method for manipulating it independently of the remainder. I consider three distinct arguments to this effect, that the components must present an affordance, that the components must be causally connected by wholly distinct mechanisms, and that the components must interact in a linear, non-additive way. I argue that this conception of modularity is mistaken insofar as it requires modularity to be an actual feature of the mechanism itself; But a close reading of **(PMa)** reveals that it is a modal notion—the conditions for its satisfaction need not exist prior to an intervention, and need not, therefore, be a feature of the mechanism itself, but could be a feature of the intervention. In other words, the conception of modularity here criticized is too narrow in focus. Thus, modularity requires only a kind of 'as-if' modular independence.

**Chapter 8** In the final chapter of the dissertation, I take **(PMa)**, the concept of as-if modular independence, and the rapprochement developed in the first half of the dissertation to demonstrate that my account, weakened though it is, can account for a range of historical cases involving the experimental manipulation of 'non-modular' mechanisms. Neither the mechanism for the action potential, nor the chemical signaling of neurons, nor the mechanisms for the perception of pain will count as modular in the sense used by critics of modularity. And, indeed, no current account of mechanistic explanation can account for these historical cases for much the same reason. Nevertheless, my account of the manipulated mechanism is able to account for these historical episodes, by giving a formal explanation for the complex interventions used in these historical episodes. I here rest my defense of the manipulated mechanism.

In this dissertation, I have motivated, formulated, and defended a quantitative account of mechanism that I call *the manipulated mechanism*. However, I have not yet shown that the manipulated mechanism can match and surpass the early successes of Woodward (2002) and Craver (2007). The manipulated mechanism holds resources for identifying explanatorily relevant components, for determining when to lump and when to split variables in a model, and for discovering the causal structure of a mechanism. Work on these research problems would form the foundation for a general account of experimental mechanism discovery that can both capture actual experimental practice in biology, and prescribe sound experimental strategies for mechanism discovery.

# Chapter 1

# Causal Models and Experimental Inference

Mechanisms are composed of parts that interact to bring about an end effect. How can we discover which components are genuine parts of a mechanism? How can we use controlled experiment to determine the contribution of a component to the end effect, and what assumptions must be made for such inference to work? Under what conditions is an explanation that appeals to a mechanism good or correct? Satisfactory answers to questions as these can only be had from a system of causal inference. This dissertation seeks to show how the tools of causal modeling—and specifically of the manipulationist program of Pearl, Woodward, and others—can be used to answer these kinds of questions. In this chapter, I introduce these tools. Chief among them are causal models, a method for representing causal networks and causal inference as graphs and graph operations, and, Woodward's flavor of manipulationism, which provide a set of assumptions and procedures for drawing causal inferences specifically from experimental evidence. In the following chapters, I will lay a foundation that shows how these tools map onto mechanisms, and how this mapping offers a way to

answer the kinds of questions above.

Causal modeling techniques offer a normative framework for causal inference, and as such are (as I will argue in Chapter 3) enormously useful for philosophical problems in mechanistic explanation. I begin the dissertation with an exploration of this normative power.

In §1.1, I present the fundaments of causal models. I introduce causal Bayes networks, systems of causally interpreted equations, and probabilistic models. I show how these three kinds of models can be brought together using bridge principles. I demonstrate briefly how causal Bayes networks have been used to solve problems in decision theory and developmental psychology.

In §1.2, I present Woodward's manipulationist account of causal relevance. Manipulationism provides a set of principles for drawing inferences on the basis of interventions into causal models. To the formal account of causal models, manipulationism adds a definition of an experimental intervention, and two constraints on the structure of causal models, *level invariance* and *modularity*. These constraints justify inferences about the structure of a particular causal system from experimental data. As such, manipulationism provides a normative framework for evaluating experimental method, and for evaluating the inferences drawn from experimental intervention. Where biologists and neuroscientists rely on experimental intervention for discovering and justifying mechanistic explanations, then Woodward's manipulationism will prove a useful tool in evaluating discovery procedures and the resulting mechanistic explanations.

## 1.1 Causal Models

Causal models are tools for describing causal structures in a very general, content-independent way. Causal models make no assumptions about the metaphysics of causation, and they do not presume anything about the subject being modeled. They are intended as a purely abstract formalism for predicting and describing causal relations in general.

The kind of causal model that I appeal to in this dissertation, called a causal Bayes network, uses probabilities do describe the relations among causal relata. The basic idea is that causal relations give rise to characteristic patterns of statistical correlation, and that we can therefore (given certain assumptions) use these patterns of statistical correlations to infer the causal structure of a mechanism.

For example, when we observe that an increase in sugar intake correlates with an increase in ATP production, we can infer that either sugar intake causes increased ATP production, or that increased ATP production increases sugar intake, or that both are the effects of an unmeasured common cause. The correlation is a sign of a causal relation connecting the two. (More information is needed to determine which of the three possibilities is the correct one, however.)

A causal Bayes net represents causal relations using a *causal graph*: A kind of diagram containing circles (nodes) that represent variables (measurable quantities) connected by arrows (directed edges) that represent direct causal relations among the variables. Thus, we could represent the claim that increased sugar intake leads to increased ATP production with the graph in Figure 1.1.

More complex causal relations, such as those in Figure 1.2, will naturally require more nodes and more edges.[1] A sequence of one or more edges is called *path*. There

[1] Although these graphs are obviously gross oversimplifications of the chemical pathways involved in energy production, they suffice to make my point.

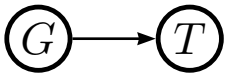


Figure 1.1: Changes in glucose intake ($G$) cause a corresponding change in ATP production ($T$).

are four paths of length one in Figure 1.2: one is a path linking $D$ and $I$ (called $DI$); the remaining three are $IT$, $GI$, and $IP$. Notice that the path $DI$ is the very same path as $ID$; the direction does not matter. There are six paths of length two: $DIT$, $DIG$, $DIP$, $GIP$, $GIT$, and $TIP$. A path over directed edges (arrows), in which each link is pointing in the same direction is called a *directed path*. A directed path represents a causal chain. There are four directed paths of length two in Figure 1.2: $DIP$, $DIT$, $GIP$, and $GIT$.

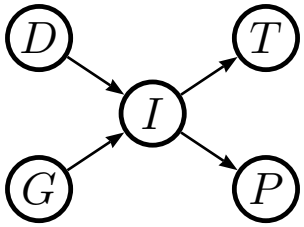


Figure 1.2: A more complex causal graph; $D$=ADP, $G$=glucose, $T$=ATP, $P$=pyruvate, and $I$=unmeasured intermediate reactions.

## Causal Bayes Nets

In a causal graph, directed edges represent direct causes, and directed paths represent chains of causation. But how can we construct a causal graph from observational data? Correlation does not imply causation, but because correlation is at least a *sign* of causation, we can use correlations along with additional assumptions, to discover

causal relations. Recall, though, that a given correlation is compatible with a range of different causal structures. The power of causal Bayes nets is that they provide a set of principles for using observed correlations to constrain the range of compatible causal graphs (ideally to just one). Bayes net modeling harnesses the probability calculus to create a metric called the screening-off condition: Correlations among two (or more) variables can be broken (or even created) when we condition on a third variable.[2]

When $X$ and $Y$ are correlated, but uncorrelated conditional on a third variable $Z$ (an operation discussed below), then $Z$ is said to screen off $X$ from $Y$. For example, let us suppose that anxiety and glaucoma are uncorrelated among the general adult population. But, as it happens, heavy coffee consumption is a cause of both anxiety and glaucoma, and so we will find that among those adults who drink large quantities of coffee, both anxiety and glaucoma *are* correlated. Now, if we condition on coffee consumption, this correlation will come apart, because there is no particular reason to think that coffee is a cause of glaucoma or *vice versa*. Coffee drinking, then, is said to screen off anxiety and glaucoma. Because common causes screen off their joint effects, screening-off is a foot in the door for inferring causation from correlation. To show how this kind of inference works, we should begin with the probability calculus.

Here are the axiomata of the probability calculus. These axiomata provide a definition of *probability*. For a set of variables $\mathbf{V}$, and for any elements of $\mathbf{V}$, $A$ and $B$:

- $P(A) \geq 0$
- $P(\mathbf{V}) = 1$
- $P(A \vee B) = P(A) + P(B)$ when $A$ and $B$ are disjoint

[2]This discussion draws from several sources: Spirtes, Glymour, & Scheines (1993); Glymour (2002); Pearl (2000); Hitchcock (2010).

These axiomata specify that the probability of any variable having a particular value or outcome will always lie between zero and one; that the probability that *some* outcome or other will occur is one; and that we can sum the probabilities of disjoint outcomes (*i.e.* a coin can come up heads or tails, but not both; thus the probability of a coin landing heads or tails is the sum of the probability it will come up heads with the probability it will come up tails).

Next, to these axiomata we add the definition of *conditional* probability:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

for $P(B) > 0$. This definition says that the probability of $A$ taking a particular value, given that $B$ is known to have taken a particular value (*i.e.* conditional on $B$) is the ratio of the probabilities of ($A$ and $B$) to the probability of $B$.

Third, we need a notion of *probabilistic independence*: We say that two variables $A$ and $B$ are independent when $P(A \wedge B) = P(A)P(B)$. Suppose that Amy and Beth each have a fair coin, and flip them. The outcomes of the two flips are independent, in that the outcome of Beth's coin flip does not in any way influence the outcome of Amy's coin flip. The probability of Amy's coin showing heads, being a fair coin, is $\frac{1}{2}$; the same holds of Beth's coin. Because the coin flips do not influence each other, they are independent, and the probability that both Amy and Beth will show heads is $\frac{1}{4}$.

But now suppose that Amy and Beth work together, and both take the bus to work, but live in different parts of town (and hence ride different routes). Ostensibly, the probability that Amy will be late is independent of the probability that Beth will be late, since they ride different routes with different conditions, *&c.* But these events are not independent because the conditions that affect their ability to arrive on time are not fully distinct: For example, both will be affected by a bus-drivers' strike, or a

traffic accident near their workplace. Conditioning on these common circumstances, however, will render their lateness independent: Once we rule out a drivers' strike, or an accident near the workplace, *&c.*, and consider only their individual circumstances, then the ability of each to make it to work on time *is* independent of the other's. In probabilistic terms, we say that two variables, $A$ and $B$, are independent *conditional* on a third, $C$, when $P(A \wedge B|C) = P(A|C)P(B|C)$.

Fourth, we assume that two variables are dependent just in case they are not independent.

Finally, we add one last assumption, a pair of bridge principles that link causation with probabilistic dependence (and hence correlation). The first is the causal Markov condition, which says that a variable is independent of its non-effects conditional on the set of all and only its direct causes. Thus, if we have some variable $A$, the set of its direct causes parents($A$)[3] , and some other variable $C$ which is not a descendant (effect) of $A$, then:

$$P(A \wedge C|\,\text{parents}(A)) = P(A|\,\text{parents}(A))\ P(C|\,\text{parents}(A))$$

which, by Bayes's theorem, is equivalent to

$$P(A|\,\text{parents}(A) \wedge C) = P(A|\,\text{parents}(A)).$$

This last form suggests one way to read the causal Markov condition: Once we know the values of all of $A$'s direct causes, learning the value of any other variable (that isn't an effect of $A$) adds nothing to our knowledge the probability of $A$.

The second bridge principle is called faithfulness (which happens to be the converse of the causal Markov condition: It says that the only conditional independencies

[3]In a directed graph, the relationships among different variables are described using kin terms. If a variables $A$ has a directed edge pointing to $B$, then $A$ is a parent of $B$, and $B$ a child of $A$. $A$ is an ancestor of $B$ when there is a directed path of any length leading from $A$ to $B$; $B$ is then $A$'s descendant. Since we are interpreting graphs causally, the parents of $B$ are all its direct causes, and the ancestors of $B$ all of its (direct and indirect) causes.

found in any graph are those implied by the causal Markov condition. In other words, it claims that the causal Markov condition won't overlook any additional conditional independencies.

The causal Markov condition permits inference from probabilistic dependence and statistical correlation to causal connection in a graph by stipulating the conditions for judging when a variable $B$ is a direct cause of another, $A$, and hence when the variables should be connected by a directed edge in the graph. The causal Markov condition says that, in a set of observational data, $B$ is a cause of $A$ when $B$ is a member of a set of variables that screens $A$ from its non-effects. There are many algorithms for using the causal Markov condition to infer causal structure from observed data, although the particulars of these algorithms are beyond the scope of this chapter.

Causal modeling using Bayes nets has seen quite wide application. One area where they have been deployed is in economic decision making. One long dominant model of decision making, evidential decision theory (Gibbard & Harper, 1981), claims that, given a decision, a range of options, and for each option a probability distribution over a range of possible outcomes, you should choose the option with greatest expected utility. The expected utility of an outcome is simply the sum of the probability of that outcome multiplied by its value; the expected utility of a choice is the sum of the expected utility over all possible outcomes for that choice.

A consequence of this theory, however, is that spurious correlations will affect the expected outcome of a decision. To borrow an example from Sober (2001), bread prices in London and Venetian sea levels are highly correlated (in that both are more or less continuously rising). We might think, if we are evaluating strategies for making bread more affordable, that one good strategy would be investing in the *Modulo Sperimentale Elettromeccanico* project, aimed at providing a kind of sea wall to protect Venice from floods. But of course this is absurd: It is absurd precisely

because there is no causal mechanism linking Venetian flooding to bread prices in London. We should, the argument goes, only act on those choices where we can have an effect on the outcome. Halting Venetian flooding has no effect on bread prices, and so we should not consider the investment as an effective option. Causal modeling is one technique for sorting out which correlations are informative for economic decision making by showing us which correlations are the result of a causal connection, and hence can be used to bring about the desired outcome (Hagmayer & Sloman, 2005; Hagmayer, Sloman *et al.*, 2007).[4]

Developmental psychology has also drawn upon causal modeling to understand how humans and other primates develop a causal understanding of the world. Psychologists have debated whether causal knowledge (or at least the ability to acquire causal knowledge) was acquired or innate. Hume famously argued that all human knowledge of the world is gleaned by the association of events or ideas; a more modern version of this idea is embodied by the Rescorla-Wagner model (Rescorla & Wagner, 1972). But more recent evidence (Miller, Barnet, & Grahame, 1995; Sloman & Lagnado, 2005) suggests that this picture is too simple. Rather, humans are capable of retroactively sorting through learned associations and developing a more refined causal picture of the world from these associations—and that this ability develops at quite an early age (Gopnik & Sobel, 2000; Gopnik, Sobel *et al.*, 2001). Gopnik, Glymour *et al.* (2004) argue that causal Bayes networks are the best foundation for a theory of causal learning in humans, by showing that human causal reasoning generally conforms to the causal Markov condition.

The causal Markov condition is bridge from observed correlations to causal models; but what about inference from experimental intervention? I turn now to consider

[4]Of course, Sober uses this example to argue that causal modeling using Bayes nets is inappropriate in these circumstances precisely because the Markov condition would recommend the absurd investment strategy described above. I am putting the example to slightly different use here.

Woodward's manipulationist account of causal modeling. Woodward offers a different set of bridge principles for moving between causal models and experimental correlations.

## 1.2 Manipulationism

Where the Markov condition permits causal inference from passive observational data, Woodward (2003) offers a method for causal inference from experimental manipulation. When we perform an ideal intervention into a causal system, and observe some variable of interest to correlate with our intervention, we can (at least oftentimes) conclude that the intervention was therefore a *cause* of the observed variable. Manipulationism is a formal rendering of the conditions and assumptions necessary to support this kind of inference. Such inference is warranted when our experimental interventions are ideal—when they are free of confounders, and are sufficiently surgical. Variables that correlate with (are probabilistically dependent on) ideal interventions are inferred to be effects of the intervention. In this way we can construct a causal graph from a systematic sequence of interventions.

### Systems of Causally Interpreted Equations

In addition to causal graphs, Woodward adds a new causal formalism: systems of causally interpreted equations. Such systems capture the same causal relationships as a graph, but also specify the *functional* relationship among the variables in the causal system. Here is a simple mechanical system composed of gears, illustrated in Figure 1.3. Attached to an input shaft $S_i$ is a gear with six teeth $G_1$. That gear turns a second gear with 10 teeth $G_2$. The second gear turns a third with 20 teeth $G_3$; the third gear is mounted on an output shaft $S_o$. We can represent this mechanism with

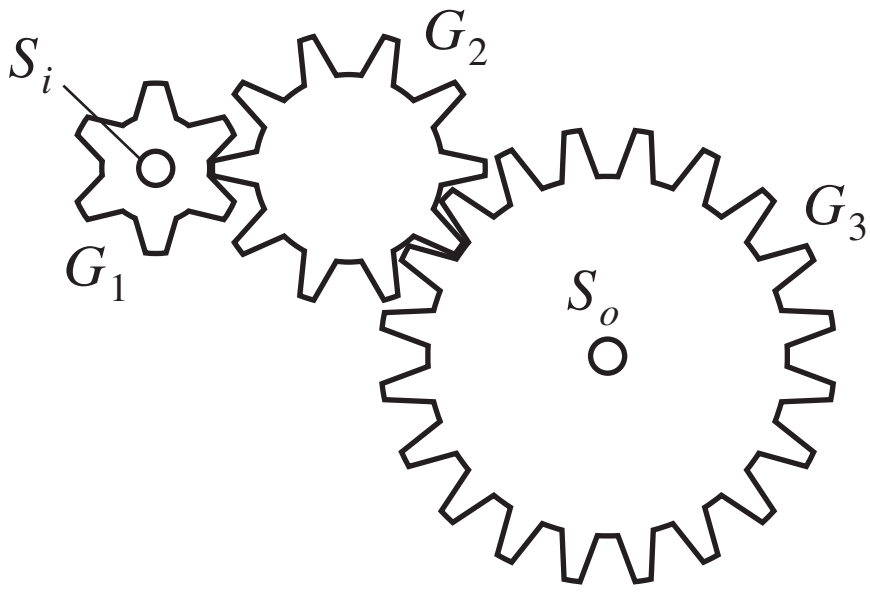


Figure 1.3: A system comprising an input shaft $S_i$ connected to an output shaft $S_o$ link by a train of three gears, $G_1$, $G_2$, and $G_3$.

a system of equations that describes the rotational speed of each gear:

$$G_1 = S_i \tag{1.1}$$

$$G_2 = \frac{6}{10} G_1 \tag{1.2}$$

$$G_3 = \frac{10}{20} G_2 \tag{1.3}$$

$$S_o = G_3 \tag{1.4}$$

In a system of equations, the '=' symbol indicates that the value of the terms to the left (the left-hand side, or LHS) is equivalent to the value of the terms on the right (the right-hand side, or RHS). In a *causally* interpreted system of equations, '=' is given the additional meaning of a causal relation, where the variables on the right-hand side are taken as the causes of the variable on the left-hand side. '=' as equivalence is symmetric and transitive. Equivalence is transitive in that since $G_1 = S_i$ and $G_2 = \frac{6}{10} G_1$, then $G_2 = \frac{6}{10} S_i$. And equivalence is symmetric in that since $G_1 = S_i$, then naturally $S_i = G_1$.

But the causal version of '=' is antisymmetric and intransitive.[5] The causal version of '=' is antisymmetric because causation is not symmetric: $G_1$'s causing $G_2$ does not alone imply that $G_2$ causes $G_1$.[6] Moreover, the causal version is intransitive because, if $G_1$ is a direct cause of $G_2$, and $G_2$ a direct cause of $G_3$, it does not automatically follow that $G_1$ is a direct cause of $G_3$ (even as it is a distal cause). We can rewrite the equation for $G_2$ as $G_2 = \frac{6}{10}S_i$, but we cannot infer from this rewriting that $S_i$ is a direct cause of $G_2$—because, in the system of gears presented here, it is not[7]

I shall say more about the features of systems of causally interpreted equations below.

## Interventions

Woodward sets out to give a semantics for causal models that captures scientific practice in giving explanations based on the results of experimental intervention (Woodward, 1997; Hausman & Woodward, 1999; Woodward, 1999, 2000; Woodward & Hitchcock, 2003a; Woodward, 2003; Woodward & Hitchcock, 2003b). The

Manipulationism claims that we can discover whether $X$ is a cause of $Y$ (and the functional relationship between the two) by *intervening into* $X$—but not just any intervention will do. Correct causal inference (Woodward argues) requires that our interventions be ideal: any measured variation in the purported effect $Y$ must be due to $X$ and $X$ alone.[8] Thus, Woodward's first axiom, *intervention*, requires that an

[5]This does not imply that such models assume causation itself is intransitive; merely that the causal operator is.

[6]I say 'alone', because of course in this example it is true that $G_2$'s moving would be a cause of $G_1$'s under certain circumstances. However, we can know this only because we have some prior understanding of how gears work; this conclusion about the symmetry between $G_1$ and $G_2$ does not follow from an algebraic manipulation of the equations. In general, algebraic manipulation does *not* preserve causal relations.

[7]But of course, depending on how the causal system is arranged, $G_1$ might well be a direct cause of the other two gears. Whether $G_1$ is a direct cause of $G_3$ is a contingent, empirical fact that cannot be derived from the algebraic manipulation of a system of causal equations.

[8]Although we must also attempt to account for measurement error and natural variation in $Y$.

ideal intervention into $X$ acts as a switch that cuts $X$ off from its other causes, and puts it under the causal control of the experimenter[9].

### Interventions in Causal Bayes Networks

Woodward's definition of an ideal intervention proceeds in two steps. First he characterizes what it is to be an *intervention variable*, for $X$ with respect to $Y$. Then, using this account of an intervention variable, he formulates the notion of intervention.

Intervention is a two-place predicate. An experiment defines a set of independent variables (the variables we will be manipulating; in general the set will be a singleton) and a set of dependent variables (the variables we will be measuring; again usually a singleton set). To simplify the discussion (without loss of generality), let us talk about single variables rather than sets. If the independent variable is $X$, and the dependent $Y$, then Woodward defines an intervention $I$ on $X$ with respect to $Y$[10] as follows:

**IV** A variable $I$ is an intervention variable for $X$ with respect to $Y$ if and only if

I1 $I$ causes $X$.

I2 $I$ acts as a switch for all the other variables that cause $X$. That is, certain values of $I$ are such that when $I$ attains those values, $X$ ceases to depend on the values of other variables that cause $X$ and instead depends only on the value taken by $I$.

I3 Any directed path from $I$ to $Y$ goes through $X$, That is, $I$ does no directly cause $Y$ and is not a cause of any causes of $Y$, if any, that are built into the $I - X - Y$ connection itself; that is, except for (a) any causes of $Y$

[9]Metaphorically speaking: Woodward allows for the possibility of 'natural interventions' which do not occur at the behest of the laboratory scientist, as well as for non-scientific interventions of the sort a small child might employ. Woodward's account is *not* anthropocentric.

[10]Although I mention three variables, intervention is actually a predicate, although as is the case here, it is often convenient to talk about it as though it were a variable. See the definition of **(IN)** below.

that are effects of $X$ (*i.e.*, variables that are causally between $X$ and $Y$) and (b) any causes of $Y$ that are between $I$ and $X$ and have no effect on $Y$ independently of $X$.

I4 $I$ is (statistically) independent of any variables $Z$ that causes $Y$ and that is on a directed path that does not go through $X$.

(Woodward, 2003, p. 98)

For an intervention to be ideal, it must be a cause of the independent variable $X$ (I1), and not merely accidentally correlated with it. We must also be sure that any change in $X$ is a direct result of our intervention *only* (I2). Since we are interested in $X$'s contribution to the dependent variable $Y$, it must not be the case that our intervention causes $Y$ directly or through any route that doesn't include $X$ (I3), else we would be measuring $I$'s contribution to $Y$, and not $X$'s contribution to $Y$. Finally, there should be no variables $Z$ that are causes of $Y$ that correlate with the intervention (I4).

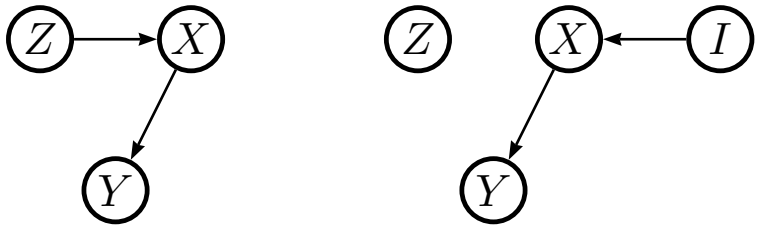


Figure 1.4: An intervention on $X$ with respect to $Y$, before and after

A simple example of an intervention variable that meets IV is presented graphically in Figure 1.4. The left side of the figure represents a causal structure over $X$, $Y$, and $Z$. We wish to intervene on $Y$ with respect to $X$, that is, we want to test whether $X$ is a cause of $Y$. The right side represents the results of that intervention. Impor-

tantly, the intervention is a cause of $X$ (satisfying I1), cuts off $Z$'s causal influence on $X$ (I2), and does not cause $Y$ directly (I3).

Once we have a variable that meets condition $IV$, Woodward defines an intervention thus:

**(IN)** $I$'s assuming some value $I = z_i$ is an intervention on $X$ with respect to $Y$ if and only if $I$ is an intervention variable for $X$ with respect to $Y$ and $I = z_i$ is an actual cause of the value taken by $X$. (Woodward, 2003, p. 98)

### Interventions in Causally Interpreted Systems of Equations

Using Pearl's (2000) notation, Woodward represents an intervention in a system of causally interpreted equations with the set operator (p. 339)[11]. We represent an intervention into $X$ to set it to some value $x$ as

$$\text{set}(X = x).$$

The set operator works by deleting from our system of equations any equation with $X$ on the LHS (with $X$ as an effect), and inserting the two new equations

$$X = aI \tag{1.5}$$

$$I = z_i \tag{1.6}$$

where $I$ meets condition IV, and $x = az_i$.

## Invariance

Not all equations can be interpreted causally. When are we warranted in giving an equation (or any generalization) a causal interpretation? Woodward's central claim is that we are so warranted when that equation is *invariant*. An invariant equation

[11] The set operator, notice, is an explicit predicate, with no mention of an intervention variable.

is one that continues to hold under (at least some range of) interventions on the variables on the RHS. When we can correctly predict the value of an effect during an intervention, we are justified in claim that that equation correctly captures the causal structure of the variables involved. Put another way, manipulating a variable should not change the way it interacts with its effects.

If $X$ is a cause of $Y$, the principle of invariance claims that an intervention on $X$ with respect to $Y$ should not disrupt the existing causal relationship between $X$ and $Y$; that is, we do not want $X$ and $Y$ to behave differently when we are intervening. If their behavior did change, we would not be able to draw an inference from the behavior of $X$ and $Y$ during the experiment to the relationship between $X$ and $Y$ under non-experimental circumstances. Figure 1.5 provides a graphical example of this concept.

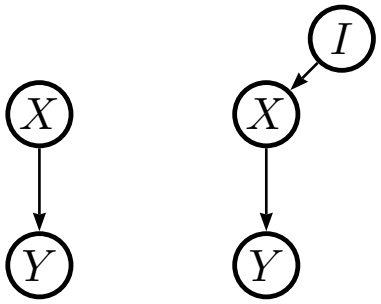


Figure 1.5: Applying the set operator on $X$, the parent of $Y$, does not change the causal relationship between $X$ and $Y$

Woodward uses a number of (*prima facie*) compatible variations on invariance throughout his book; I have chosen the last and most general formulation for this discussion. Woodward axiomatizes invariance as Probabilistic Level-Invariance:

**PLI** $Pr(X|\,\text{parents}(X)) = Pr(X|\,\text{set}(\text{parents}(X)))$. (Woodward, 2003, p. 340)

This can be read to say that intervening into the causes of $X$ does not disrupt the relationship between the causes of $X$ and $X$ itself. If, to take the automobile example, I were to experiment by manipulating the gas pedal and observing the resulting velocity, I would not expect that merely pressing on the gas pedal would alter the nature of the relationship between pedal position and velocity change.

## Modularity

The third and final principle in Woodward's system is the modularity condition. Where invariance is concerned with interventions into single equations, modularity extends this concern to interventions on systems of equations; modularity requires that interventions do not disrupt the system as a whole. Simply put, intervening on a variable should not disrupt the other causal relations in our model. When causal relations are expressed as equations, an intervention on one equation should leave the other unrelated equations unchanged and invariant. If, in a system, $X$ causes $Y$, and $A$ causes $B$ (and that's all), then interventions on $X$ or $Y$ should not disrupt the relationship between $A$ and $B$. Figure 1.6 demonstrates this principle graphically.

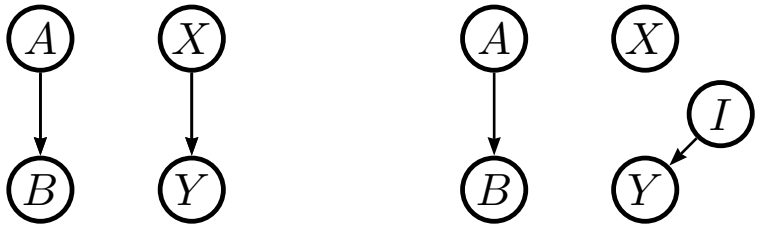


Figure 1.6: Interventions on $X$ or $Y$ have no effect on the relationship between $A$ and $B$.

The precise specification of the modularity requirement has undergone a number of revisions in response to critics. Hausman & Woodward (1999) provide the earliest

specification of interest here. This formulation was later updated by Woodward (2003) in response to criticism (Cartwright, 2002). Although I return to consider this issue in Chapters 5, 6 and 7, consider here a few attempts at formulating the principle:

**(MOD)** A system of equations is *modular* if (i) each equation is invariant under some range of interventions and (ii) for each equation there is a possible intervention on the dependent variables that changes only that equation while the other equations in the system remain unchanged and invariant. (Woodward, 2003, p. 329)

Additionally, Hausman & Woodward show that **(MOD)** implies a weaker but closely related claim, **(MOD*)**.

**(MOD*)** For all distinct variables $X$ and $Y$, if $X$ does not cause $Y$, then $Pr(Y \& \text{set}(X)) = Pr(Y) \cdot Pr(\text{set}(X))$. (Hausman & Woodward, 1999, p. 553)

Woodward also provides a probabilistic restatement of **(MOD)**, designed as a bridge principle for inferring causal structure from dependencies observed during an experimental intervention.

**Probabilistic Modularity (PM)** For all distinct variables $X$ and $Y$, $Pr(X | \text{parents}(X)) = Pr(X | \text{parents}(X) \& \text{set}(Y))$, or equivalently, $X \perp \text{set}(Y) | \text{parents}(X)$. (Woodward, 2003, p. 340)

Once one has conditioned on the causes of a variable, $X$, setting any distinct variable makes no difference to the probability distribution of $X$. That is, interventions on $X$'s parents do not disrupt the relationship between $X$ and parents($X$) (the invariance condition) and interventions on variables other than $X$'s parents either make no difference to $X$ (if, say, the intervention is on one of $X$'s children) or are screened off by parents($X$).

These three principles—intervention, invariance, and modularity—provide the conditions under which we can bridge a set of probabilistic dependencies observed during an experimental intervention with a causal graph to create a complete causal model. The set of tools Woodward gives are designed to mirror the way experiments are conducted and interpreted in the social sciences and elsewhere, but have been carefully regimented to give the account a normative bite.

Manipulationism's normative bite comes from the constraints that the principles of intervention, invariance, and modularity place on experimental interventions. Manipulationism claims that, when these principles are satisfied, we are entitled to draw certain causal inferences from an experimental intervention. And when these principles are not satisfied, that we do not have such entitlement. As such, manipulationism places constraints on what counts as a *good* experimental design,[12] and hence on what counts as a *good* causal inference from experiment.

## 1.3 Conclusion

In this chapter, I have reviewed the components of a causal model—causal Bayes networks, probabilistic dependencies, and causally interpreted systems of equations—, and principles that allow us to bridge them. Causal Bayes nets represent causal relations as circle-and-arrow diagrams; probabilistic models represent correlations as dependency relations, although they do not encode causal relations directly. The causal Markov condition and the faithfulness condition are assumptions that permit us to pick out which of the dependencies in a probabilistic model represent causal relations, and hence permit us to construct causal Bayes networks on the basis of observed probability distributions. Causally interpreted systems of equations represent

[12]Not all experiments involve interventions, of course, and manipulationism is silent about those experiments that do not.

causal relations as (possibly non-deterministic) mathematical functions.

Woodward's manipulationism provides a semantics for interpreting a causal Bayes network as a model for predicting the outcomes of experimental interventions. Woodward's concepts of an intervention, and his principles of invariance and modularity provide bridge principles for constructing causal Bayes networks on the basis of probability distributions observed under the conditions of experimental intervention. Invariance and modularity are also principles of causal inference: They are rules for constructing and updating graphical models in the face of experimental evidence.

Early accounts of mechanistic explanation and mechanisms face a cluster of problems centered around a common thread: How can we determine whether a particular component is a part of a mechanism? Biologists usually turn to experimentation to answer this question. Here, then, is where the principles of manipulationism hold interest for the philosopher of biology: In guiding the inference from experiment to mechanistic structure.

I turn now to introduce mechanistic explanation. In the chapter following, I show how the causal modeling concepts introduced in this chapter can be applied to some of the philosophical problems that mechanistic explanation faces.

# Chapter 2

# Qualitative Accounts of Mechanism

In the previous chapter, I introduced causal modeling and manipulationism. Together, these provide a powerful epistemological framework for drawing causal inferences from experimental data. In this chapter, I give an historical overview of mechanistic explanation, to highlight the steps taken. I argue that the next natural move is to show how mechanisms can be represented using causal models, and how these models facilitate causal inference about mechanisms.

Mechanisms are, to put it very roughly, a set of components that interact so as to produce an end effect. Biologists, in particular, take their job to be, in part, the description of the mechanisms in living nature that produce the phenomenon they seek to explain.

But, we might ask, what justifies a biologist's claims about the workings of a mechanism? These claims are usually derived from a causal interpretation given to some set of experimental data. As yet, our best accounts of mechanistic explanation have only begun to make progress at showing how these inferences can be justified.

In this chapter, I trace the recent history of mechanistic philosophy, and show how accounts have risen to meet the challenge of accounting for scientists' behavior.

I will flesh out the problem of experimental inference to mechanisms, and show that meeting it is now very salient to mechanistic philosophy. In the following chapter, I will show how the most recent accounts of mechanism have taken the first steps at meeting this problem, and that there remains significant work to be done. Finally, in Chapter 4, I return to the tools of causal modeling, to show how they can be applied to solve this problem.

In the final chapter of his *Four Decades of Scientific Explanation* (1989), Salmon raises several challenges for philosophers of science. Chief among them is this: That a complete ontic[1] account of causal explanation must face (what I call) Hume's Question.

**Hume's Question** What is the secret connexion that links cause to effect?

On Salmon's ontic view, to explain an event is to exhibit the causes of that event: to situate the event in an etiology. But unless we can give a convincing account of what it is that binds cause and effect, this kind of explanation is too shallow to do any real scientific work. Explanation differs from mere description in that explanations can answer a range of 'what if things had been different?' questions. Without an account of causation at its heart, ontic explanation would be incapable of answering these kind of counterfactual questions. For those that shared Salmon's ontic view of explanation, finding a suitable response to Hume was of utmost importance.

Glennan (1992, 1996, 2002), sharing Salmon's ontic view, sought an answer to Hume's Question. Looking to how the sciences treat causation, Glennan suggests

[1] as opposed to a purely formal account, *e.g.* the DN model of Hempel & Oppenheim (1948).

that Hume's secret connexion is a *mechanism*, in the most general sense. In §2.1, I review Glennan's earlier (1996) account of mechanism, and his view of how an account of mechanism can answer Hume (returning to his later views later in this chapter, and in the next chapter). Glennan first account was ultimately, however, unable to provide a satisfactory account of the secret connexion, because, as he admits, there is some bottom-most level of basic physics for which we will not be able to analyze the causal relations into mechanisms. And, at this level, Hume's Question remains unanswered.

Contemporaneously with Glennan, Bechtel & Richardson (1993) began a parallel project. They did not seek to answer Hume's Question; their interest was descriptive: They sought to make sense of scientific explanatory practice, and in particular, the methods by which scientists discovery mechanisms. Noticing that many scientific explanations appeal to mechanisms, Bechtel & Richardson set out a descriptive project aimed at characterizing the heuristics deployed to construct these kind of explanations. In §2.2 I lay out their characterization of mechanisms as comprising parts that are *decomposable* and *localizable*. Bechtel & Richardson show that the assumptions of decomposability and localizability, while often false, have proven useful heuristics for the discovery of mechanisms. These features will resurface in later treatments of mechanistic explanation.

Bechtel & Richardson's (1993) mechanistic project sought to accurately describe scientific practice; however, other authors wanted to bring their descriptive concept of mechanism to bear directly on problems facing accounts of scientific explanation. Machamer, Darden, & Craver (2000) take Glennan's causal notion of mechanism and Bechtel and Richardson's notions of decomposability and localizability to craft an account of mechanistic explanation in fulfillment of Salmon's ontic view. In §2.3, I review this account, examining their precise formulation of mechanism in some detail,

as it is the best developed view of mechanism.

As Machamer, Darden, & Craver are quick, and right, to criticize Glennan's (1996) reliance on causal laws, Glennan (2002) offered a refinement of his account that replaced talk of causal laws with talk instead of Woodward-style invariant generalizations (see §1.2). Glennan's ultimate aim was to show that mechanisms provide the truth-makers for manipulationist counter-factual claims. Psillos (2004), however, offers a strong argument that Glennan has put the cart before the horse: Counterfactuals, Psillos argues, are more basic than mechanisms, and any future account of mechanism had better recognize this. Psillos's argument sets the stage for the move towards a rapprochement of manipulationism and mechanism, and therefore warrants close examination in §2.4.

## 2.1 Glennan's Mechanism

I use Glennan's (1992) PhD dissertation and the resultant publications (Glennan, 1996, 1997a,b) as a starting point for thinking about the mechanistic movement in the philosophy of science. Although earlier authors (*e.g.* Salmon, 1984) certainly made use of mechanistic concepts in their writing, Glennan's achievement was to give one of the first clear analyses of the concept of mechanism. Although Glennan has since offered significant alterations to this account, his early work is, I believe, a good place to start a discussion of mechanisms generally.

Glennan was less interested in *scientific* explanation than he was in answering Hume's Question. Hume asks: What is the secret connexion that links cause and effect? Glennan responds: A mechanism. Glennan argues that we can distinguish genuine causes from spurious correlation by looking for a mechanism that links a purported cause to its effects. Thus, on Glennan's view, when we explain $E$ by appeal

to its cause $C$, our explanation is incomplete until we specify the mechanism that underlies the causal relation from $C$ to $E$.

So, what is a mechanism? Glennan gives this account:

**(M)** A mechanism underlying a behavior is a complex system which produces that behavior by the interaction of a number of parts according to direct causal laws. (Glennan, 1996, p. 52)

Notice that **(M)** defines mechanism with respect to an 'underlying a behavior'. Since Glennan is interested in the connection of cause to effect (reading, perhaps contentiously, 'behavior' as 'effect'; I will have more to say in defense of this identification in the next chapter). Why should a mechanism be defined relative to some behavior or effect? A given physical system, Glennan observes, will exhibit indefinitely many behaviors (has indefinitely many effects) (Kauffman, 1970). And, in any system, there will be an indefinite number of ways to decompose it into potential mechanisms. By defining a mechanism relative to a single behavior, we can be sure that there is (in a given explanatory context) one and only one correct way to decompose a given complex system. In this way, Glennan avoids any worries that the concept of mechanism doesn't lead to subjectivism about causal explanation: Whatever our explanatory interests may be, once we have fixed those interests, there is an objective fact of the matter as to which specification of a mechanism will satisfy those interests.

Glennan gives two features of this account further elaboration: *law-governed interactions* and *parts*.

## Laws

The interactions among the parts of a mechanism, according to **(M)**, are governed by *direct causal* laws. A law is causal if it relates cause and effect; a law is direct if

it requires no intermediaries between cause and effect. To explain what he means, and to motivate this proposal, Glennan considers two cases. First, although there is a true counterfactual supporting the generalization that night follows day, such a generalization is not *causal*, because it describes the correlation between two effects that share a common cause (the rotation of the earth). Second, consider a system of three gears, connected in a sequence. Although there is a true causal generalization that describes the motion of the third gear in terms of the motion of the first, it is not *direct*, because the generalization leaves off the intermediate gear, which is the direct cause of the third gear's motion. In contrast, a generalization that describes the third gear's motion in terms of the second would be a direct causal law, because there are no intermediaries between the second and third gear that are left off. [2] By requiring that the laws be direct causal laws, Glennan attempts to capture the sense that a mechanism's behavior "stems from a series of local interactions between parts" (Glennan, 1996, p. 56).

Glennan takes a broad view of laws, following Goodman (1947): "Laws are generalizations (or universal propositions) which support counterfactuals. Law-like or nomic generalizations are distinguished from accidental generalizations because accidental generalizations offer no such support," and laws are true law-like generalizations (Glennan, 1996, pp. 54–55). Glennan takes it as a virtue of this account that it makes no distinction between 'deep' or fundamental laws, and higher-level counterfactual supporting generalizations. This lack of distinction is a virtue for Glennan, because it allows him to articulate *e.g.* social mechanisms, whose parts are governed by generalizations that fall short of full-blown lawhood.

[2] Directness is ambiguous in the context of a real mechanism, of course: Why is the second gear the direct cause of the third's rotation, and not, say, tooth no. 35, or the forward-facing surface of tooth no. 35, or so on? Glennan suggests that we think about directness as relative to a given set of parts. If our decomposition of the gear mechanism bottoms out at gears, then the second gear is a direct cause. If our analysis bottoms out at teeth (and other gear-parts), then tooth no. 35 is the direct cause.

However, as Glennan notes, the concept of counterfactual support is usually taken to be a *causal* concept. But if counterfactual support is part of the analysis of mechanisms, and mechanism is part of the analysis of causality, and causality a part of the analysis of counterfactuals, then a reliance on counterfactual support renders Glennan's account circular. Glennan has a way out of this circularity that I will explore below.

### Invariant Generalizations

But **(M)**'s dependence on direct causal laws, Glennan realized, was fraught with difficulty. Laws are problematic entities for philosophy of science generally, but too, many scientific disciplines, *e.g.* biology and medicine, do not seem to even *have* laws (Machamer, Darden, & Craver, 2000).

Woodward (2003) argues that the distinction between universal laws and accidental or contingent generalizations is ill-founded—precisely because the special sciences lean very heavily on a class of such contingent generalizations. Woodward offers an alternative. He says, instead, that the important generalizations for science are *invariant*—generalizations that hold up under experimental test. (See §1.2) We can distinguish causal relations from mere correlations by testing them, and observing whether the relation in invariant under intervention. The observed correlation, for example, between thunderstorms and barometer fallings, does not hold when we intervene into the barometer to set it to a particular value: Thunderstorms form independently of any such manipulation.

Thus, Glennan (2002) (looking ahead slightly) offered a modification to **(M)**[3] **(M′)**:

[3]I depart from Glennan's usage, by naming his modification as **(M′)**, where Glennan continues to call this new principle by the old name.

**(M′)** A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations.

## Parts

Mechanisms, Glennan claims, are decomposable into parts, and a good mechanistic explanation therefore explains by exhibiting how the parts interact to produce the mechanism's behavior. The idea that mechanisms are composed of parts is a significant claim, and one that following accounts adopt as a defining characteristic of mechanisms. But what is a part?

If we restricted our notion of parts to just physico-mechanical parts, Glennan's mechanisms would be nothing more than what he calls mere *machines*—a distinction that Glennan is very careful to draw (Glennan, 1996, p. 51). Rather, he desires that his account should extend beyond mere machines: He wants electrical circuits (which have no moving parts), chemical pathways involved in gene expression (which appeal to genes—problematic entities in many ways, and spatially diffuse chemical interactions), and computer software (which have no physical parts at all) to count as mechanisms too. But if anything of our choosing can count as a part, then Glennan's account is rendered vacuous—there would be no constraint on how to decompose a complex system, and so any decomposition would count as a mechanism. Thus Glennan is forced to constrain what can count as a part: Parts are robust, and have an independent existence from the mechanism as a whole. "It should in principle be possible," he tells us, "to take the part out of the mechanism and consider its properties in another context" (Glennan, 1996, p. 53).[4] In short, parts must be, as he says,

[4]To foreshadow slightly, we see in this requirement the beginnings of a modularity requirement for mechanisms.

"objects" (p. 53).

The important consequence for this restriction on what can count is a part is that it generates conditions for determining when mechanistic explanation 'bottoms-out'. Glennan notes that a part $p$ of a mechanism need not be a simple, that is, $p$ could itself be decomposable as a mechanism. But if part $p$ is *not* itself decomposable into further parts, because $p$ is not composed of *objects* which have independent existence, then the mechanistic explanation has bottomed out—we can conclude that $p$ is a simple and neither has nor needs any further explanation.

Glennan offers the electromagnetic field as an example of a simple part. The electromagnetic field is, on Glennan's account, an object—it can play the role of a part in a mechanism. We can decompose the electromagnetic field into two vector fields, the electric field $\vec{E}$ and the magnetic field $\vec{B}$. Having done this, Glennan notices that we can attribute to any particular point in space an electric field strength and a magnetic field strength. However, Glennan argues that such points in space cannot be *parts* of a mechanism for the electromagnetic fields, because points in space are not suitably independent. For example, although we can choose the strength of the *e.g.* electric field at one point of space, we cannot do so independently of the field strength at adjacent points in space. Moreover, points in space cannot be bounded, or excised from the electromagnetic field. Indeed, points in space are not objects in any sense of the term. So, although we can conceptually decompose the electromagnetic field in this way, such a decomposition is not a *mechanistic* decomposition. Thus, Glennan concludes, there is no the mechanism for electromagnetic field, and we can therefore treat it as a simple part.

Glennan takes this as a significant result for his account, because the electromagnetic field is what he calls a law-governed entity, the kind of entity that fundamental physics deals in, and that eludes causal explanation—because there are no causes at

the fundamental level.[5] Thus, his account has a principled way of correctly identifying law-governed entities: Law-governed entities are all and only those entities for which we cannot give mechanistic explanations.

## Circularity, Recursive Mechanisms, and Hume's Puzzle

Thus fully elaborated, let us put **(M)** to Hume's Question. We make a claim: $C$ explains $E$ because $C$ causes $E$. Hume asks: What is the secret connexion between $C$ and $E$? Glennan responds: There is a mechanism, composed of parts $\mathbf{P}$ (where, I presume, $C \in \mathbf{P}$[6]) that interact according to direct causal laws. Thus, in $P$ we have an causal-mechanical explanation of the causal relation between $C$ and $E$.

But now, Glennan notes, Hume can re-raise his Question about any of the lawful interactions among the members of $\mathbf{P}$. Suppose that as part of the mechanism, $P_1$ causes $P_2$, in accordance with some direct causal law $L$ that links the two. Hume will once against ask: What is the secret connexion between $P_1$ and $P_2$?

This re-raising of Hume's Question is what I call Hume's Puzzle. Like a petulant child continually asking 'why?', Hume can always re-raise his Question.

> It is confessed, that the utmost effort of human reason is to reduce the principles, productive of natural phenomena, to a greater simplicity, and to resolve the many particular effects into a few general causes... But as to the causes of these general causes, we should in vain attempt their

[5] I should be very quick to add that Glennan hangs nothing on the term 'governed'; he uses the term only as a colloquial short-hand for a part which cannot be decomposed further, and whose behavior is completely captured by a fundamental law of physics.

[6] It is worth noting that neither Glennan nor Craver think that mechanisms are always etiological, with the explanandum behavior manifested in some causally-downstream part. Rather, they permit a constitutive view of mechanisms, allowing that the behavior of a mechanism may be a complex description of large swaths (or even the entire) mechanism, and may involve both $C$ and $E$. So, in this sense, my characterization of Glennan here is unfair. Nevertheless, my goal here is only to provide a simple example that demonstrates one of the strengths of Glennan's view, even as such ignores an important nuance. I will have more to say on constitutive vs. etiological mechanisms in the next chapter.

> discovery; nor shall we ever be able to satisfy ourselves, by any particular explication of them.... The most perfect philosophy of the natural kind only staves off our ignorance a little longer. (Hume, 1777, p. 26)

A mechanistic account of causation—as Glennan well recognized—must not only answer Hume's Question, but Hume's Puzzle as well.

Glennan's proposed solution to Hume's Puzzle is simple and elegant. Returning to the question of the connection between parts $P_1$ and $P_2$, Glennan says either $P_1$ and $P_2$ are connected by a mechanism—that is, the direct causal law $L$ that links the parts is a mechanically explicable law—, or they are are not connected by any mechanism—that is, $L$ is a fundamental law. Mechanisms, in short, are finitely recursive entities. A mechanistically explicable law is a law that can be explained by appeal to a sub-mechanism. A fundamental law goes hand-in-hand with Glennan called 'law-governed entities'[7]—they are a brute part of fundamental physics, for which no response to Hume's Question can be given. Thus, each time Hume re-raises his Question, we have a ready answer. Either there is a sub-mechanism involved (to which Hume can raise his Question yet again), or we can claim to have bottomed out.

This recursive nature of mechanisms is Glennan's response to the problem of circularity raised by analyzing laws causally. If we ask for an explanation of $C$ causing $E$, and we appealed to a mechanism that relies on laws that (in effect) claimed merely that $C$ was a cause of $E$, we would have a problem of circularity. But Glennan's recursive mechanisms do not work like that: We can explain $C$'s causing $E$ by appeal to a mechanism that relies on causal laws—but these laws will in turn appeal to the causal workings of sub-mechanisms within this mechanism. And so our explanation is *not* circular.

[7]Again, nothing hangs on the term 'governed'; Glennan only uses it as a perhaps unfortunate but somewhat clearer stand in for 'part whose behavior is completely capture by a fundamental law of physics, and not by a further mechanistic decomposition'.

But, Glennan worries, when we hit the end of the recursion, we bottom out at mechanically *in*explicable causes at the fundamental level. What is our response to Hume now? Do we just shrug out shoulders? Do we deny the existence of causation at the level of fundamental physics? And, either way, does this move void the mechanical account? We might worry, as Glennan do, that shrugging our shoulders when we bottom out means that (M) is but a fancy gloss on a regularity theory of causation.

Glennan argues that his mechanistic account is much more than a simple regularity account dressed up in fancy clothes. Glennan shields mechanisms by distinguishing between two kinds of cause. He admits that "[a]t the level of fundamental physics, Hume's problem still remains" (Glennan, 1996, p. 68)—there are regularities that we call causal, but for which no explanation is forthcoming, no answer to Hume's Question. But at higher levels, causation can be explicated mechanically and non-circularly, and that this mechanical notion of causation is autonomous from the fundamental, inexplicable notion. That is, even while mechanical explanation takes advantage of a decomposition of the explanandum phenomenon, it does not reduce the explanandum phenomenon to the resultant components: "we can distinguish between connections and conjunctions, because we can understand the mechanisms which produce higher level regularities" (Glennan, 1996, p. 68). That is, mechanistic explanations at higher levels are autonomous, without need for appeal to lower levels.

Nevertheless, because there is no account of the secret connexion for fundamental laws, Hume's Question will, at some point in this recursive process, have no answer. And so long as Hume is unsatisfied, we have work to do in putting together a complete account of mechanistic explanation.

## 2.2 Discovering Mechanisms

Whereas Glennan took himself to be solving Hume's Problem, Bechtel & Richardson were concerned with making sense of how scientists discover mechanistic explanations. When they started the work a decade earlier (Bechtel & Richardson, 1993, p. xi), they found in their case studies that, *contra* Hempel & Oppenheim (1948) (and *contra* Glennan (1996), but *cf.* Glennan's later revisions to eliminate laws from his account (2002), discussed below.), scientists in biology, psychology, and medicine rarely, if ever, cited laws in their explanations.[8] Indeed, there was nothing in their explanations that resembled D-N explanation at all. Rather than cite arguments, or discuss the expectability of an explanandum, Bechtel & Richardson discovered that scientists largely prefer to explain by laying out the *mechanism* by which the explanandum is brought about.

On Bechtel & Richardson's view, mechanistic explanation describes a kind of machine (in a very broad sense) responsible for bringing about an explanandum phenomenon. Thus, like Glennan, Bechtel & Richardson take mechanisms to be defined relative to some phenomenon; there are no mechanisms *simpliciter*.

> **(B&R)** A [*mechanism*][9] is a composite of interrelated parts, each performing its own functions, that are combined in such a way that each contributes to producing a behavior of the system. A *mechanistic explanation* identifies these parts and their organization, showing how the behavior of the machines is a consequence of the parts and their organization. [emphasis added] (p. 17)

Although Bechtel & Richardson do not provide any additional detail, one can easily

[8] Indeed, one may well wonder what a law of biology or a law of psychology would even look like. *Cf.* Machamer, Darden, & Craver (2000)

[9] Bechtel & Richardson use here 'machine', but intend the same meaning as where other authors use 'mechanism'. I have edited the quote for consistency with Glennan's (1996, pp. 51–2) usage of 'machine' and 'mechanism'.

see their commitment to the ontic conception of explanation (or, at least, that they attribute this commitment to the scientists they present). An explanation, on their view, is a text that refers to the mechanism in the world—and nothing more. They are nonetheless unconcerned about answering Hume's Question.

Bechtel & Richardson were primarily interested in how mechanistic explanations were discovered and developed over time. Thus, in contrast to Glennan, who focused on the technical apparatus of **(M)**—a necessary step to answering Hume's Question—Bechtel & Richardson do not dwell on (what I call) **(B&R)**, but immediately identify a pair of heuristics that scientists frequently use when constructing a mechanistic explanation.

Mechanisms, note Bechtel & Richardson, are composed of parts. And thus, if we want to construct a mechanistic explanation, we must identify the parts of the mechanism that figure in that explanation. Bechtel & Richardson observe that life scientists have often turned to the heuristics of *decomposition* and *localization* to discover the parts of a mechanism, and hence to construct mechanistic explanations.

## Decomposition

The scientists that Bechtel & Richardson present worked with complex phenomena. Gall sought to explain general understanding, which encompasses large swaths of human cognitive achievement of varying kinds; Pasteur sought to explain fermentation, which encompasses a wide range of different chemical interactions. *Decomposition* involves breaking up such a complex phenomenal behavior into distinct and simpler sub-behaviors, each of which is easier to investigate:

> Decomposition assumes that one activity of a whole system is the product of a set of subordinate functions performed in the system. It assumes that

> there are but a small number of such functions... [that] are minimally interactive... [and] can be handled additively or perhaps linearly. (Bechtel & Richardson, 1993, p. 23)

Notice what decomposition presumes of a mechanism: That the parts of a mechanism interact in a (mostly) additive fashion. Of course, many (if not most?) complex phenomenon are not the additive resultant of the behaviors of parts, as Bechtel & Richardson well realize; nevertheless, as a *heuristic*, decomposition can get complex projects rolling, even if the strategy does often ultimately fail.

## Localization

Decomposition is a conceptual task; the sub-behaviors identified in decomposition must then be identified within the complex system responsible for the complex phenomenon behavior. *Localization* is the process of identifying parts or functions of parts within a system as producing the sub-behaviors obtained from decomposition. Of course, behaviors cannot often be localized in this way, even when a decomposition succeeds. Sometimes a sub-behavior is diffuse within a system. For example, against Gall, Flourens found that the understanding was not localizable to any particular part of the human brain proper, but was diffuse across all parts of the brain. Still, as a heuristic, Bechtel & Richardson maintain that localization is a fruitful strategy.

Although decomposition and localization will often fail, they provide a "*tractable* strategy for attacking problems" (p. 27, emphasis mine) if not for solving them. This is important because, say Bechtel & Richardson, most of the complex phenomena that interest life sciences are really too complex for human minds to get a grip on, and as scientists are (currently) human, the heuristics provide a method for getting a cognitive foothold into explaining such complex phenomena.

These two accounts see two distinct roles for mechanistic explanation to play. Whereas Glennan was concerned with crossing one of the last looming hurdles for causal explanation—answering Hume's Question—, Bechtel and Richardson's project was more historical than philosophical. They were concerned with showing both that scientists really do rely on mechanistic explanation, and in showing how such explanations are discovered. One might take Glennan's project as *metaphysical*: He was concerned with answering Hume's Question, of showing how a causal link can be decomposed into a mechanism; whereas one might take Bechtel and Richardson's project as *descriptive*: They were concerned with showing how scientists actually *do* decompose a causal link into a mechanism. Machamer, Darden, & Craver (2000) attempted to synthesize these two projects, to create an account of mechanistic explanation.

## 2.3 MDC

Machamer, Darden, & Craver (2000), take these two views of mechanisms due to Glennan and Bechtel & Richardson, and synthesize them into a new account, not of causation, but of mechanistic explanation. Their essay—I will follow common parlance in referring to it as *MDC*—brought wide attention to the otherwise previously obscure mechanisms literature. Like Bechtel and Richardson, MDC observed that life scientists describe mechanisms in their explanations; if our philosophical accounts of explanation are going to make any sort of contact with actual scientific practice, then we had better have an account of mechanistic explanation. Like Glennan, MDC understood that giving an ontic account of mechanistic explanation required solving Hume's Problem. However, unlike Glennan, MDC did not offer mechanisms themselves as a response to Hume; instead, they appealed to *activities* as the secret connexion.

MDC were moved to write, as on their view, "there is no adequate analysis of what mechanisms are and how they work in science" (p. 4). The analyses of Glennan (1992, 1996), and Bechtel & Richardson (1993) fall short of adequacy, as on MDC's view, they both failed to emphasize the importance of *activities*.

The emphasis of these earlier analyses, MDC note, had been on the component *parts* of mechanisms and their interactions. Glennan emphasizes the *lawful interactions* of parts; Bechtel and Richardson emphasize the *contribution* each part makes to the explanandum behavior. But although being composed of interacting parts is certainly a key feature of a mechanism, it is not the only feature.

Instead, MDC emphasize that when the parts of a mechanism interact, these interactions are *productive*, not just of the final explanandum behavior, but of changes within the mechanism itself. 'State changes' or 'interactions' are thin concepts that cannot capture *how* those changes come about, or "the productivity by which those changes are effected" (p. 5). MDC looked for a thicker concept, one that includes the productive nature of these interactions, and one that gives some sense not just that change occurs, but the concrete, particular nature of that change. This concept, MDC claim, is not simply the properties and property transitions that entities have and undergo, but is something more: A productivity—although the precise nature of that productivity is left unexplored.[10] MDC call these productive relations *activities*.

Too, MDC note that Glennan's (1992; 1996) account of mechanism, perhaps the best contender for an adequate analysis, leans heavily on the idea that the interactions are the result of direct causal laws. Laws, MDC observe, rarely figure in molecular biological or neuroscientific mechanistic explanations. Thus, a proper account of mechanisms will replace a problematic reliance on laws with a reliance on a the thicker notion of activities.

[10]But see the Anscombian project of Bogen (2008), which aims at adding flesh to this very concept.

Thus, Machamer, Darden, & Craver offer the following characterization of mechanisms. They claim that

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

The notion of *entities* in **(MDC)** is little different than the notion of parts in **(M)** or **(B&R)**—namely, that entities (or parts) be spatiotemporally located bearers of properties, which are capable of engaging in activities. While **(MDC)** does not explicitly *require* entities to be *localizable* (in the sense of Bechtel and Richardson), it nevertheless implies that entities typically will be.[11] Many of MDC's mechanism examples are composed of localizable parts: DNA replication (p. 3), chemical neurotransmission (pp. 8–ff.), DNA transcription (pp. 15–ff.). Tellingly, when describing neurotransmission they use several two-dimensional diagrams that spatially and temporally locate the discrete parts of this mechanism, and note that "mechanisms are often represented this way." (p. 8). Moreover, MDC claim that "[t]raditionally one identifies and individuates entities in terms of their properties and spatiotemporal location"—which is a hedge, to be sure, but a strong hint that they would prefer parts to be localizable.[12]

The notions of *activities* and *regularity* are new additions to the analysis, and so demand further attention.

[11] Contrast *located* with *localizable*. Intelligence—that generic cognitive faculty—is clearly located somewhere, namely in a subject's head. But notice that being able to locate intelligence does not imply that it is localizable: Intelligence (if there is any such thing) may be diffuse across the entire cortex of the brain, and be heavily intertwined with many other cognitive functions.

[12] Although Craver admits to the hedge (in personal correspondence), he does observe that "they have to be somewhere".

## Activities

What makes the concept of activities a richer concept than simple interactions or state changes? Activities are types of causes, MDC say, but scientists invoking mechanisms do not typically use such abstract terms as 'cause'. Instead, scientists invoke richer, more specific verbs and gerunds. Following Anscombe (1971), MDC observe that the word 'cause' "only becomes meaningful when filled out by other, more specific, causal verbs, e.g., scrape, push, dry, carry, eat, burn, knock over" (p. 6), and that a component is only counted as a cause (in the abstract sense) once it has been identified as engaging in some specific activity. Being a cause is parasitic on engaging in an activity.

But simply being a richer concept does not mandate the inclusion of activities in an analysis of mechanism. Although MDC have claimed that scientists make use of these richer concepts, and that the concept of causation is simply an abstraction from the myriad activities that entities do engage in, no explicit argument is made for their inclusion in **(MDC)** above and beyond a simple appeal to causes.

One natural interpretation of the move to include activities is that activities are meant as a response to Hume's Puzzle. Recall that Glennan responded to Hume's Puzzle with a finitely recursive appeal to mechanisms. Another possible response to Hume's Puzzle is to nip it at the bud: To give a response to the first instance of Hume's Question that does not permit it being re-raised. This is how empiricists as Norton (2003) respond: By claiming that there is no necessary connection, and therefore no causation (and no Question to be raised). Such a response is definitive, and so solves Hume's Puzzle.

MDC take a similar strategy (but *sans* the causal denial), by offering a definitive response to the first instance of Hume's Question. What is the connection between cause and effect? In this instance, it is pushing; in another, electro-motive force; in

yet another, combustion. These are all activities—all causal notions that, because they are thick concepts, provide an answer to Hume's Question. Activities are the kind of thing that can be discovered by empirical investigation. It is an empirical fact that $C$ is related to $E$ by the activity of combustion. And so in this sense, not only do activities answer Hume's Question, they answer it definitively, preempting the re-raising and so solving Hume's Puzzle, because there is no further need for investigation: We can see the secret connexion—activities—for ourselves.

## Regularity

Although **(M)** makes no explicit mention of regularity, Glennan clearly intended that mechanisms were the kind of thing whose behavior was stable over time. Indeed, this is one of the supposed virtues of laws, is that they explain the regularity of the phenomena they subsume, and Glennan does, recall, invoke direct causal laws in **(M)**. That mechanisms act in accordance with such laws explains their regularity.

However, MDC eschew the notion of direct causal laws. First, while the subjects of biology and neuroscience obviously exhibit regularities, such regularities are not laws in the usual sense—they are contingent, hold often only for the most part, they often vary substantially from one kind of organism to another. Laws, on the other hand, are usually taken to be necessary, universal, and with no reference to particulars (see, *e.g.*, Hempel & Oppenheim, 1948).

But without a reliance on laws MDC must add the further stipulation that mechanisms are regular "in that they work always or for the most part in the same way under the same conditions" (p. 3). Crucially missing from **(MDC)** and its discussion, however, is any account of a mechanism's regularity. MDC claim that "these regularities are non-accidental and support counterfactuals to the extent that they describe activities" (p. 7),—the presence of activities is clearly *meant* to explain the

regular behavior of a mechanism, but no story is given for how activities are to play that role. Activities might be on-off, non-deterministic, or even random, for example. Although activities have an intuitive appeal as a source of regularity, a principled account of how this works is still required.

Still, after the publication of MDC, one might well wonder what philosophical work an appeal to mechanism is supposed to do. Glennan thought a proper account of mechanism would solve Hume's Problem; by now it is clear that this cannot be. MDC thought that activities could solve Hume's Problem, but, as I will show, Psillos (2004) made it clear that this cannot be either.

If mechanisms cannot solve Hume's Question, then what does the mechanistic mode of explanation offer above and beyond competing theories of causal explanation, such as the physical causal theories of Salmon (1994) and Dowe (2000)? These accounts of physical causal explanation at least purport to have solutions to Hume's Problem without the mechanistic apparatus, which should make them appealing alternatives. Or consider the manipulationist account developed by Woodward (1997, 1999, 2000, and his conspirators). Manipulationism purports to have an account of causal explanation that doesn't require a response to Hume's Question in the first place. Thus: Where do mechanisms fit in? What problem are they to solve?

One obvious strength of mechanistic explanation is that they have a descriptive adequacy lacking in physical causation accounts, and only nascent in Woodward's manipulationism. Thus, mechanistic explanations bring us closer to actual scientific practice, rather than forcing us to rely on abstractions or reconstructions of scientific reasoning.

Psillos (2004) finds another equally important role for mechanistic explanation. In raising a challenge to activities, he finds an important niche for mechanisms to fill: They can tell us which counterfactual claims are also causal claims, by showing

that there is a mechanism responsible for the truth of the counterfactual. That is, the descriptive richness of mechanism is a necessary component for a fully fleshed-out counterfactual explanation. I turn now to consider Psillos's arguments, both against and for mechanism, as his view opens the door to a quantitative account of mechanism to complement the descriptive strengths of mechanistic explanation.

## 2.4 Psillos's Challenge to Activities and Counterfactuals

Recall that Glennan offered **(M)** as a response to Hume's Question. Indeed, as Psillos notes, Glennan took **(M)** as "an unproblematic way to understand the counterfactuals which [mechanisms] sustain" (Glennan, 1996, p. 63). When we say that, had $C$ been otherwise, so too would have $E$ (because $C$ causes $E$), Glennan says that the truth of this counterfactual claim is that there is an **(M)** mechanism that links $C$ to $E$: That mechanism "explains why the counterfactual holds" (Psillos, 2004, p. 308). Mechanisms are the truth-makers for these kind of counterfactual claims.

Glennan (2002), recall, offered a new way to understand mechanisms that replaced a reliance on direct causal laws (which, MDC and even Glennan himself worried, were mysterious and, at any rate, play at most a small role in biology) with a reliance of invariant generalizations—a particular class of counterfactual-supporting generalizations relating cause and effect that is stable under interventions. But now Psillos worries that **(M′)** (**(M)** revised in this way), introduces a circularity: Mechanisms explain counterfactuals, but counterfactuals explain the workings of a mechanism. Although Psillos does not make the further claim that this circularity is *vicious*, the circularity should certainly give one pause.

Moreover, Glennan distinguishes between *mechanically explicable laws* and *funda-*

*mental laws*. A mechanically explicable law is one that can be explained by mechanisms; a fundamental law cannot. Fundamental laws occur in the most fundamental physics. Glennan cites Maxwell's equations governing the electro-magnetic field as an example of a fundamental law: There is no mechanism for electro-magnetic wave propagation. Nevertheless, Psillos notes (p. 309 ff.), such laws clearly support counterfactuals. So **(M)** mechanisms must not be a necessary condition for explaining counterfactuals: There must be a further story about the truth of counterfactuals to be told. Indeed, since any particular mechanism will bottom-out in such fundamental laws, Psillos wonders whether mechanisms can play the role of truth-maker for counterfactuals. Indeed, Psillos does not deny that mechanisms can *explain* mechanically explicable laws—but they cannot be part of the truth conditions for those laws, as they are rendered redundant by the fundamental laws upon which they supervene.

Consider an example: I claim that, were I to push the pedal, this automobile would accelerate. We can explain this counterfactual claim by an appeal to the various bits and bobs contained the pedal, fuel injector, and so forth. But, Psillos argues, these bits and bobs, even as they offer a constitutive explanation of the acceleration, are not themselves part of the truth conditions for that counterfactual: The truth of the counterfactual is in the fundamental details of molecular motion and electro-weak interactions upon which the story about the bits and bobs depends.

Thus Psillos concludes, at least for Glennan's view, because fundamental laws or mechanistically inexplicable invariant generalizations support counterfactual claims, counterfactuals are more fundamental than mechanisms—mechanisms cannot, *contra* Glennan, explain counterfactuals. Rather, Psillos concludes, it must be the other way around.

## Activities are a Red Herring

Psillos argues that the same holds of MDC: They too put the mechanistic cart before the counterfactual horse.

Machamer, Darden, & Craver, recall, include *activities* in their ontology. They view entities and activities as distinct ontological categories, on a par. The properties of an entity constrain what activities it may engage in; likewise, activities constrain what entities may participate in those acts. "Entities and activities," Machamer, Darden, and Craver say, "are correlatives. They are interdependent" (p. 6). One cannot exist without the other (p. 8).

On the MDC view, there is a symmetry in the interdependence between activities and entities. But, Psillos points out, there is good reason to think that the relationship between activities and entities is asymmetrical. First, (p. 312) Psillos notes that (acknowledging one of MDC's points) although we cannot have activities without entities, we *can* have entities without activities—my glasses, for example, sitting on my desk beside me, are in no way acting. Activities do not seem to have the same independent existence that entities enjoy. Can there be throwing without an object thrown? Psillos does not see how an activity can determine the kinds of entities that can so act. The activity of throwing does not determine what can be thrown; rather it is having a certain size and mass that determines whether an entity can be thrown. These, size and mass, are properties of a particular entity, not properties of the activity. Once you have the entities, you have the activities; they come along for free. So, Psillos concludes, activities are redundant.

But even if activities weren't redundant, they do not avoid counterfactuals. Activities were posited, recall, as a response to Hume's Question. MDC could not give an account of mechanistic explanation without an account of causation. But Psillos does not think that activities are a suitable response, because they cannot avoid

counterfactuals. Consider the mechanism for the action potential in neurons. At one stage, a change in the potential across the neuron membrane *triggers* voltage gated ion channels to open. Triggering is an activity that the potential field and the ion channels participate in. But what does triggering mean? Psillos would argue that it means something like this: If the potential *hadn't* changed, the ion channel would not have opened. Activities require counterfactuals in order for us to make sense of them. Thus, positing activities does nothing to respond to Hume's Question, because they do not and cannot explain the counterfactuals they allude to. Thus, Psillos concludes, activities do not explain counterfactuals. Rather, it must be the other way around.

## Counterfactuals are the Key

Psillos concludes that neither can the Glennan nor the MDC account of causation eliminate counterfactuals from causal explanations; Mechanisms are not the truth-makers for causal counterfactual claims. Glennan's account uses mechanisms to explain counterfactuals, but then turns to a different set of counterfactuals to explain mechanisms, which in itself is not problematic. The problem is that this recursion bottoms out, not in mechanisms, but in counterfactual claims that have no mechanistic explanation. The MDC account relies on activities to explain counterfactuals, but activities themselves only make sense if we make a further appeal to yet more counterfactual claims.

One might read this conclusion as dismal; if mechanisms rely, ultimately, on counterfactuals, then we might worry that counterfactuals are doing the explanatory heavy lifting in a mechanistic explanation. What role, then, do the mechanisms themselves play in the explanation (aside from giving the explanation a narrative that matches scientific practice in describing causal explanations)?

Nevertheless, Psillos was optimistic that mechanisms could support explanations

in non-experimental contexts, where the truth of causal counterfactual claims could not be verified experimentally. For example, in history, one might establish causal links by pointing to a plausible mechanism that appears to link cause and effect. But this optimism, I think, is a bit deflationary, because it presumes that mechanism talk is just giving a nice narrative gloss over an empirically defeasible observation of a chain of cause and effect. This deflation misses something important about mechanisms.

Mechanisms are more than a story, because not just *any* causal system will count as a mechanism. There are constraints on which causal systems can count as a mechanism. Indeed, precisely because mechanisms rely on counterfactual claims—claims that can be experimentally tested—, these constraints are (as I will show in the next chapter) powerful tools for guiding experimental inference about mechanisms. So, although Psillos's dismissal of mechanisms to the realm of informed speculation undervalues mechanistic talk, his negative conclusion in fact points the way forward: We can harness the counterfactuals that underlie mechanisms as an inferential tool.

## Responses to the Counterfactualist View

Glennan (2011) has responded that mechanisms and counterfactuals are not asymmetric. Although, he admits, mechanisms do depend on counterfactuals, counterfactuals also depend on mechanisms: "what makes a certain counterfactual claim true," Glennan says, "is that there is a mechanism that would respond in a certain way to a manipulation" (Glennan, 2011, p. 24). Where Glennan reads Psillos as drawing a reductionist conclusion—namely, that mechanisms are reducible to counterfactuals—Glennan here argues that the dependence of mechanisms on counterfactuals does not entail the further claim that mechanisms *reduce* to counterfactuals, because counterfactuals (or at least causal counterfactuals) depend on mechanisms for their truthmakers. Thus, the dependence between mechanisms and counterfactuals is mutual,

and there can be therefore no reduction of one to the other.

On Salmon's view, what explains are things in the world: The mechanism itself explains the causal link, not a description of the mechanism. On this ontic view of explanation, if, as Psillos clearly wants, counterfactuals are to have any explanatory import it is because they (or their truth-makers) are a feature of the explaining mechanism. However, Bogen has argued for two claims that block Psillos's move. First, Bogen (2004) argues that there are no actual truth-makers that can underwrite counterfactual claims in any kind of non-trivial way[13] Bogen (2005, 2008) makes the further argument that counterfactuals are not themselves a feature of causal links, because causation does not reduce to any kind of regularity, including the kind of regularity implied by causal counterfactual claims. So, causal links (and by extension mechanisms) exhibit neither the truth-makers for counterfactual claims, nor the regularities implied by them. Bogen departs from Psillos in that he disputes that any coherent account of activities must rely on counterfactual claims. Rather, he argues, the claim that one thing caused another need only rest on the actual facts surrounding the two occurrences, and that these facts are sufficient for causal explanations. Thus, counterfactuals are not a proper part of (ontic) causal explanation. Instead, he argues that instead of general, abstract accounts of causation, we should look to Anscombian activities to underwrite our causal explanations. This move has the added benefit of matching actual scientific practice, which, he notes, is generally devoid of overt counterfactual claims.

Nevertheless Bogen does not dispute that counterfactuals play an important *epistemological* role. We do use counterfactual reasoning to infer the existence of causal relations, even if the counterfactuals are not constitutive of those relations. It is a category mistake to say that this kind of inference works because of some modal feature

[13]Trivially, all counterfactual claims are true, because counterfactuals are a species of conditional in which the antecedent is necessarily false. This kind of truth is uninformative and uninteresting.

shared by all causal relations; yet, it is true that causal inference does seem to require induction over counterfactual claims, because causal inference relies, in large part, on making predictions: We say, if this thing is a cause of that, then were this to wiggle, so would that. Such a prediction is readily testable, by making the counterfactual consequent factual, and evaluating the antecedent causal claim appropriately.

I think that it is important to observe that both sides of this debate allow that counterfactual reasoning is central to the discovery of mechanisms. I will return to develop this point in greater detail in Chapters 5 and 8 where I argue that, *contra* Bogen, counterfactual reasoning about the causal structure of mechanisms does in fact require that mechanisms exhibit a kind of modal feature.

## 2.5 Conclusion

Mechanistic explanation is a kind of causal explanation that takes a cause and an effect, and demonstrates how the cause brought the effect about by showing that the causal relation is brought about by a number of interacting parts. Since mechanisms comprise causal relations among their parts, and causal relations comprise mechanisms, mechanisms can contain sub-mechanisms, or can be parts in super-mechanisms. The causal structure of the world can thus be hierarchically organized. However, mechanisms must bottom-out in counterfactuals; below a certain point, the causal relations in a mechanism are taken as brute, with no further explanation of the counterfactuals that those bottom-most relations support. This bottoming-out may be conventional (molecular biology, for example, bottoms out at the level of molecular and atomic mechanisms, and molecular biological explanations as a rule do not invoke, *e.g.* quantum-level mechanisms), or they may be physical (fundamental physics is engaged in the very search for that which is brute in the physical world; that which

*cannot* be explained by further decomposition).

Although mechanistic explanation nicely captures scientific practice in giving explanations, they fall short as metaphysical accounts of causation. Mechanisms, Psillos (2004) has argued, do not explain counterfactuals; rather, counterfactuals explain mechanisms. If so, one might well wonder what role mechanistic explanation has to play, if all of the heavy explanatory lifting is being borne by, say, manipulationist accounts of causal explanation.

Yet, nothing in Psillos's arguments undermines mechanistic explanation's descriptive strength. Indeed, his negative thesis sets the stage for a novel mechanistic thesis: What if manipulationism could be harnessed to provide a quantitative account of mechanism to complement the qualitative accounts glossed here? Manipulationism provides an account of the counterfactuals that hold of a mechanism's parts, where the mechanistic accounts provide constraints on what kinds of causal structures are to count as genuine mechanism, constraints that closely match scientific practice in identifying mechanisms. If such is possible, we would have the tools to show not just *that*, but *how* counterfactuals underlie the mechanistic explanations biologists give, and how these counterfactuals could be harnessed to discover mechanisms experimentally, and to evaluate the resulting mechanistic explanations. Manipulationism opens the door to a quantitative and a *normative* account of mechanistic explanation.

Woodward (2002) and Craver (2007) have seen this opportunity, and have begun a project of crafting a quantitative account mechanism. In the next chapter, I turn to consider the motivation for rapprochement in more detail, and examine the promise of these two projects. I also consider what must be done in order to complete these projects, to craft a fully quantitative account of mechanism with sufficient resources to also provide a normative account of mechanistic explanation.

# Chapter 3

# Quantitative Accounts of Mechanistic Explanation

In this chapter, I consider the normative turn in the mechanistic literature. The descriptive accounts presented in the previous chapter do not, generally, attempt to provide a framework for evaluating mechanistic explanations according to a set of criteria. More recently, however, both Woodward (2002) and Craver (2007) have taken steps forward in developing a normative account of mechanistic explanation, accounts that would have the resources to evaluate mechanistic explanations by bringing the formal apparatus of manipulationism to rapprochement with the descriptive accounts of mechanism. I will argue that although both accounts represent concrete steps towards rapprochement, neither quite gets us there. Woodward's account looses sight of the descriptive accounts of mechanism; Craver's account looses sight of the formal apparatus.

In the previous chapter I introduced mechanistic explanation. In this chapter, I consider how we might go about evaluating a mechanistic explanation. Some mechanistic explanations are better than others; Craver (2007) distinguishes between complete and incomplete explanations (mechanism sketches), and between more concrete and more abstract explanation (mechanism schema). We might think, for example, that the best mechanistic explanations are complete and maximally concrete. We might also think that a good mechanistic explanation will leave out irrelevant details, deliberately tacit background conditions, and spuriously correlated phenomena. A *normative* account of mechanistic explanation is one that provides a clear set of desiderata, and principles for judging mechanistic explanations against these desiderata.

I have two goals in this chapter. First, I set up and defend the need for a normative account of mechanistic explanation—one that has the resources for evaluating a mechanistic explanation against a reasonable set of criteria. Second, I set out and evaluate two recent steps towards at creating a normative account of mechanistic explanation.

In §3.1 I introduce the need for a normative account of mechanistic explanation, argue that a rapprochement with the formal apparatus of manipulationism is a fruitful approach to developing such an account.

In §§3.2–3.3 I examine two recent steps toward a rapprochement made by Woodward (2002) and Craver (2007). Although both accounts recognize the value of bringing manipulationism to mechanism, I will argue that both leave us with open questions about how this rapprochement is best accomplished. In particular, neither attempt quite succeeds in formulating a principle of mechanistic relevance that balances the need for descriptive adequacy against the formal components of manipulationism.

Woodward's (2002) account, which I consider in §3.2, analyzes mechanisms purely

in terms of the patterns of counterfactual dependence associated with manipulation. He analyzes a mechanism (or rather, a representation of a mechanism) as a set of entities, and a set of generalizations invariant under interventions that describe the relations of counterfactual dependence among those entities. Mechanistic relevance, on this view, just is causal relevance. But, as I will argue, Woodward's analysis does an injustice to the qualitative accounts of mechanism it purports to account for: not everything that is causally relevant is mechanistically relevant (indeed, this was part of the very motivation for developing an account of mechanistic explanation in the first place).

In contrast, Craver's (2007) account, which I consider in §3.3, analyzes mechanisms in terms of part-whole relations. Craver is primarily interested in developing an account of mechanistic explanatory relevance based on Woodward's manipulationism. Craver observes that we can manipulate a whole by manipulating its parts; as can we manipulate the parts by manipulating the whole. Only genuine parts will exhibit this behavior, and thus we can discover whether a part is relevant by applying the test of *mutual manipulability*. But, as I will argue, constitutive relevance is not a relation that can be adequately captured with the formal apparatus of manipulationism. In particular, there is no clear way to represent interlevel relationships using causal graphs, nor any clear way to represent interlevel interventions.

In the next chapter, I will offer my own analysis of mechanistic relevance that bridges the descriptive with the formal as the basis for a normative account of mechanistic explanation.

## 3.1 The Need for Normativity

We have already accounts that do an excellent (if imperfect) job of accounting for how biologists give mechanistic explanations. What further need is there for an account that can evaluate the explanations given, and the process by which these explanations are derived? I see two broad reasons why we should want to take this extra normative step, one external and one internal. The external reason is that biologists themselves have implicit epistemological norms for constructing and evaluating explanations, standards that are fair game for philosophy; the internal reason is that meeting the broader philosophical challenges of explanation requires an account of mechanistic relevance. Manipulationism, I will argue, is a fruitful starting point for constructing a normative account of explanation because it can address both of these projects simultaneously.

### Discovery and Evaluation

Darden (2006) argues that the process of mechanism discovery—one of the primary aims of biology—is one of continual construction, evaluation, and refinement. Her interest is in collecting and generalizing the strategies for construction, evaluation, and refinement that biologists use in the process of discovering mechanisms. But, Darden is careful to point out that her task is not one of developing norms for how mechanistic explanations *should* be constructed, evaluated, and refined. The strategies she develops "are 'advisory,' not descriptive or prescriptive" (p272). "In a future discovery episode," she continues, "the philosopher may be able to provide advice that one or more of these strategies may prove useful."

But we are left wondering: How can we know these strategies are useful? That they worked in the past is no guarantee that they will work in the future. And if

they are useful, we might wonder what features of these strategies make them useful? Under what circumstances will they be useful and why? The future biologists we are advising will surely ask these questions—they will, rightly, want to know that our advice is sound advice.

Philosophers have long studied epistemology and the norms of inference. There is nothing special about scientific reasoning that puts it out of bounds, not even that biologists are generally good making inferences about mechanisms. If we really want to say something about strategies for construction, evaluation, and refinement of mechanistic explanation, we need to draw out scientists' hidden assumptions and inferential practices. If our goal is the best science possible, it seems incumbent upon philosophy to do just that.

At times, Darden even seems to agree. Drawing upon Craver's (2007) distinctions between how-possibly, how-plausibly, and how-actually mechanisms, and mechanisms sketches, schemata, and full explanations, Darden has set up a system for evaluating mechanistic explanations. A more concrete mechanistic explanation is better than a more abstract explanation; a mechanistic explanation with fewer gaps is better than one with more gaps; a mechanistic explanation that conforms to empirical constraints is better than one that is merely possible (pp. 289–ff). All this seems to suggest that Darden is very much engaged in a normative project, because she has an account of better and worse mechanistic explanations.

But these norms, as Darden admits (even as she does not admit that they *are* norms), are just a first step (pp. 306–306). A fully-fleshed out how-actually mechanistic explanation might yet not be very good, if, for example, it contains a great many irrelevant factors, or gets the order of events wrong, a point I turn to now.

## Explanatory Relevance

Hempel & Oppenheim's (1948) DN model of explanation faced a wide range of problems from within philosophy (and without). The DN model was plagued by problems of relevance (Kyburg, 1965; Salmon, 1979), asymmetry (Bromberger, 1966), making it the hard way (Suppes, 1970), and of accidental generalizations (Salmon, 1998). These problems, described below, are ongoing concerns for any account of explanation, and so our best accounts of mechanistic explanation had better be able to handle them.

Craver (2007) has argued that a successful account of scientific explanation must be able to offer delineation principles along (minimally) each of these axes:

**Relevance** In an explanatory text, some elements will be relevant to the explanation, and some will not. Good explanations include all and only the relevant elements.

**Asymmetry** The relationship of explanation links a set of explanans (which do the explaining) to the explanandum (which is to be explained). This relationship is typically not reversible—the tides do not explain the orbit of the moon. Good explanations get the ordering correct.

**Making it the Hard Way** . The explanans need not render the explanandum *expectable* in any sense. Sometimes the explanans can render the explanandum unlikely (as when a pool player makes a very difficult shot). Good explanations do not rely on expectability in all cases.

**Accidental Generalizations** . No clear conception of laws is forthcoming; nevertheless, some generalizations are more explanatory than others. Good explanations rely only on explanatory generalizations.

Thus, a successful account of scientific explanation will:

1. Discriminate the explanatorily relevant from the explanatorily irrelevant,

2. Correctly identify the direction of explanation,

3. Not conflate expectability with explanation, and

4. Discriminate explanatory generalizations from non-explanatory.

These desiderata stem from our tutored philosophical intuitions about what makes for a good explanation. Salmon (1989) took these desiderata as pointing to a causal theory of explanation: Causes are explanatorily relevant to their effects; the direction of explanation is the direction of causation; effects need not be expected from their causes; and causal generalizations are explanatory where non-causal or accidental generalizations are not.

But it is not clear that the early descriptive accounts of mechanistic explanation are up to the task, especially with respect to relevance. None of the accounts discussed in the previous chapter offers a sense of what it means for a component to be mechanistically explanatorily relevant, aside that it must be a part of the mechanism. But how can we determine which components are mechanism parts? Only once we have a principled answer to this question can we begin to satisfy Craver's desiderata.

Glennan (2002); Woodward (2002); Craver (2007) have all independently identified Woodward's manipulationism as a framework that could be used to craft an account of mechanistic explanation that has an answer to the mechanistic relevance question, and hence would satisfy all four desiderata. Each of these accounts ask that we adopt Salmon's view that causation is the key to all four desiderata. Then, adopting Woodward's manipulationism, each account characterizes the causal relationship between two entities as a relation of counter-factual dependence under interventions. In this way, manipulationism appears to hold the tools for satisfying Craver's desiderata by bringing Salmon's causal relevance to mechanistic explanation.

Indeed, a rapprochement of mechanism and manipulationism should help us with the external challenges, discussed above. Darden (2006) argues that many of the strategies for evaluation and refinement hinge on experimental evidence, but does not offer a clear vision of how experimental evidence should alter our mechanistic hypotheses. Manipulationism holds that causal relevance is closely bound up with experimental manipulation. From an initial standpoint at least, manipulationism looks fruitful for thinking about this inferential relationship. I will argue in following chapters that, in fact, it is.

## Moving Forward with Rapprochement

However, application of the manipulationist framework to mechanistic explanation is not a straight-forward affair, because mechanistic explanation, as presented in the previous chapter, is a slightly different animal than the kind of causal explanation that Woodward was after. Causal relations are central to mechanisms, yes, but where a Woodwardian explanation would be content to stop there, mechanistic explanation requires more. For one, many (though by no means all) of the mechanists are dissatisfied with an account of causation that rests on relations of manipulability; the causal relations in mechanisms are built from productive activities (Bogen, 2004; Machamer, 2004), which are not merely counterfactuals or regularities. Moreover, if we restrict ourselves to causal relations, we miss the unique structural features of mechanisms: Mechanisms are hierarchical, comprising sub-mechanisms; and they are etiological, linking a cause with an effect.

Thus, we should ask ourselves: How can we apply the tools of manipulationism to specifically *mechanistic* explanation, and yet do justice to these central features? Too, as Darden is at pains to point out, biologists practice does seem to work for them, and she is right to defend the idea that they do not need 'saving'; thus, our application of

manipulationist tools should not run roughshod over actual scientific practice either. We should ask, how can we use these tools to lay manifest the assumptions and implicit inferences biologists make?

I turn now to survey two recent steps forward in answering this question, two attempts at rapprochement between manipulationism and mechanism.

## 3.2 Woodward's Quantitative Mechanism

Woodward (2002), unsatisfied that then-current accounts of mechanism could expand beyond the life sciences, uses manipulationism as the basis for a generalized account of mechanism. As with *MDC*, Woodward faults Glennan (1996)[1] for his reliance on direct causal laws—not only are laws not to be found in, *e.g.* biology, they often aren't even found in physics. Given that mechanisms are structured entities whose components are described with causal generalizations (and not laws specifically), Woodward is concerned with how to construct a causal model of a mechanism that adequately captures the mechanism's structure. He presents a set of necessary conditions for a causal model to be specifically a model of a mechanism.

### (MECH)

Woodward begins by considering a block sliding down an inclined plane (Woodward, 2002, p. S367–ff.), as in Figure 3.1. In the ideal case, the effect of gravity on the block's acceleration is given by the equation

$$a = g \sin\theta - \mu_k g \cos\theta \tag{3.1}$$

[1]He was, I am supposing, unaware at the time of Glennan's (2002) concession to invariant generalizations—as both papers were presented at the very same symposium.

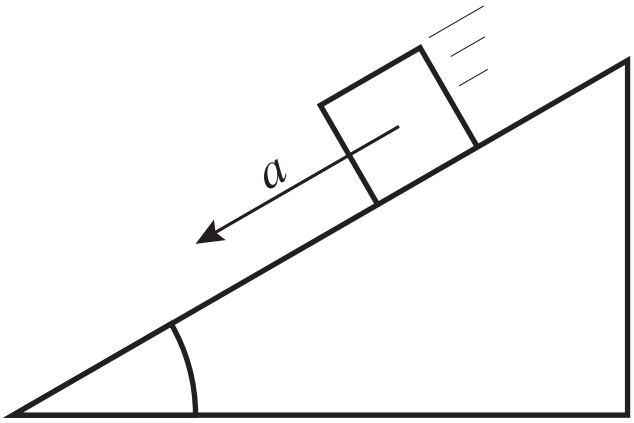


Figure 3.1: A block sliding with acceleration $a$ down a plane inclined to angle $\Theta$. Adapted from Woodward 2002, Figure 1, p. S376.

where $a$ is the magnitude of the block's acceleration, $g$ the gravitational constant of $9.8\text{ms}^2$, $\mu_k$ the coefficient of dynamic friction, and $\theta$ the angle of the incline (p. S368; with corrections).

Equation (3.1) is not (Woodward claims) a law, even while it describes a productive (i.e. causal) relationship. The equation is limited in scope, holds only approximately, and only holds at all within a narrow region of space (that is, near the surface of the earth) (p. S368). But it does correctly describe how, for example, altering the angle of the incline or greasing the slope would change the acceleration experienced by the block. That is, it correctly describes the causal relationship between gravity, friction, and angle on the movement of the block. It is an invariant generalization, in the manipulationist sense of the term.

Woodward, agreeing with Machamer, Darden, & Craver (2000) and Glennan (2002), claims that mechanisms, such as the mechanism for the acceleration of the block, should be construed as relying on, not laws, but what he calls *invariant generalizations*.[2] This concept, he claims, does not suffer the drawbacks of an appeal to

[2]Although Machamer (2004) has individually rejected manipulationism as anything more than an epistemic tool, and Darden has shied from making claims either way, it seems to me that the

laws—there are invariant generalizations in biology, invariant generalizations need not be necessary, unlimited in scope, or even widely applicable to correctly underwrite our causal claims. They need only be such that they correctly predict the consequences of (some) interventions into the causes—the variables on the RHS. (For a more detailed treatment of causal interpretation of equations, refer back to Chapter 1.)

Woodward considers too how mechanism parts fit with invariant generalizations. He argues that that *modularity* constrains the possible decompositions of a mechanism. The modularity condition, recall from §1.2, requires that in a causal model of more than one equation, that interventions into one equation not disrupt the remaining equations. Put slightly differently, if an intervention into one equation *does* disrupt the remaining equations, then there is likely a problem with the model, specifically, that there is some causal relationship involving the intervened-into variable that is incorrectly represented in the model (or not represented at all). But how does modularity constraint the decomposition of a mechanism?

The components of a mechanism each engage in one or more distinct behaviors, each of which, Woodward assumes, can be captured with an invariant generalization. The interactions or activities within a mechanism are the result of the components engaging in these behaviors. Components of a mechanism are distinct insofar as they engage in distinct behaviors. Removing or disrupting a component will remove the behaviors it contributes from the working of mechanism as well. Likewise, removing a behavior from a mechanism can only be achieved by removing the component responsible for it. The removal of a behavior from a mechanism is captured by the

account of activities presented in Machamer, Darden, & Craver (2000) does at least begin to point down this path (*cf.* Psillos, 2004). They say that activities support regularities that are "non-accidental and support counterfactuals to the extent that they describe activities. For example, if this single base in DNA were changed and the protein synthesis mechanism operated as usual, then the protein produced would have an active site that binds more tightly" (pp. 7–8). While not explicitly manipulationist in character, this passage suggests that their conception of activities is nevertheless amenable to the manipulationist view.

elimination of the invariant generalizations that describe that behavior from the representation. Modularity requires that, if when we remove or disrupt one behavior, a second behavior is removed or disrupted as a result, those behaviors *must* belong to the same component. Thus, we can know if we have decomposed a mechanism correctly when our model comprises distinct behaviors. And we can know if our model comprises distinct behaviors because the generalizations will conform to modularity.

From the notions of invariant generalizations and modularity, Woodward constructs this characterization of a mechanistic model:

**(MECH)** A necessary condition for a representation to be an acceptable model of a mechanism is that the representation

(i) describe an organized or structured set of parts or components, where

(ii) the behavior of each component is described by a generalization that is invariant under interventions and where

(iii) the generalizations governing each component are also independently changeable, and where

(iv) the representation allows us to see how, in virtue of (i), (ii) and (iii), the overall output of the mechanism will vary under manipulation of the input to each component and changes in the components themselves.

(Woodward, 2002, p.S375)

(i) simply restates the basic idea of a mechanistic explanation due to, *e.g.*, Machamer, Darden, & Craver (2000); (ii) is where Woodward connects activities with invariant generalizations; (iii) and (iv) together comprise a statement of the modularity condition described above, which governs how a mechanism can be decomposed into parts.

**(MECH)** is specifically designed to bridge the qualitative analyses of mechanism with Woodward's manipulationist framework, and hence stands as a qualitative analysis of mechanism. A graphical model of a mechanism would use, **(MECH)** tells us, variables to stand for components, and directed edges to stand for activities, as Woodward (explicitly, but perhaps unreflectively) links components with variables, and activities with edges. Invariant generalizations, recall, represent causal connections; the causal relata are the components themselves. Thus do activities *qua* invariant generalizations correspond to edges and components to variables.

But this is the full extent of Woodward's attempt to harmonize his account with the forgoing qualitative accounts. **(MECH)** falls flat as an attempt to bridge the qualitative and quantitative, as it runs roughshod over several important aspects of the qualitative accounts of mechanism. First, although Glennan and MDC are careful to point out that mechanisms are always mechanisms *for* some behavior, **(MECH)** places no such constraint on mechanisms, nor does it offer a principle by which to determine which components are or are not relevant to the mechanism. Second, mechanisms are posited to explain known (or hypothesized) cause-and-effect relations; **(MECH)** makes no attempt to link a putative cause to an effect, or to offer constraints on how to model a mechanism that does. I turn now to examine these shortcomings in detail.

## (MECH) and Mechanism Bounding

My first objection to **(MECH)** arises from the fact that it fails to offer a bounding principle. Notice that Woodward has eschewed Glennan's idea that there are no mechanisms *simpliciter*: Where Glennan argued that mechanisms are defined with reference to an explanandum phenomenon, Woodward makes no such requirement. MDC are adamant that mechanisms have start and stop conditions, but such a con-

straint makes no appearance in **(MECH)**. The only constraint that **(MECH)** places on the structure of a mechanism is that it contain component parts and their behaviors. **(MECH)** in particular permits any component to participate in a mechanism, so long as it is causally relevant. Similarly does **(MECH)** permit us to arbitrarily exclude any component as a member of a mechanism.

One might defend Woodward at this point by observing that **(MECH)** only characterizes *representations* of mechanisms, and not mechanisms themselves; that therefore it is not incumbent on **(MECH)** to lay out any membership or other structural constraints as these are well-handled by the qualitative accounts themselves.

Such a defense is disingenuous, however, as neither do Glennan's account nor MDC's account provide such principles of membership; indeed, many had hoped that the formal apparatus of manipulationism might supply such principles. The onus should be on the quantitative accounts to supply these principles. In which case, we should prefer to **(MECH)** a quantitative account that can add such a principle to the extant descriptive accounts.

## (MECH) and Mechanistic Explanation

My second objection to **(MECH)** arises from the lack of bounding principles. With no bounding principles, Woodward does not constrain mechanisms to lie between an explanandum cause and an explanandum effect. Thus, **(MECH)** does not require a mechanism to link an explanandum cause to an explanandum effect—Woodward has identified mechanisms with causal structure *simpliciter*, and leaves for mechanisms, therefore, no explanatory work to do. The explanatory heavy lifting, on Woodward's (2003) view, is borne entirely by invariant generalizations.

A mechanistic explanation is a response to a request to explain how it is that a cause and an effect are so linked. We may observe that $A$ causes $B$, and then ask:

What explains the link between the two? The response, on the mechanistic view, is a description of the mechanism that links $A$ to $B$. But for Woodward, what explains the link between $A$ and $B$ is a relationship of counter-factual dependence that is invariant under intervention—that we can manipulate $B$ by intervening to change $A$ explains the causal link between $A$ and $B$.

Thus, for Woodward, an invariant generalization that describes the relationship between $A$ and $B$ is (minimally) explanatory, where for the mechanist, it is not explanatory at all (because generalizations are not mechanisms).[3] There remains a fundamental disconnect between **(MECH)** and the descriptive accounts of mechanism about what is doing the explanatory heavy lifting, a disconnect that must be bridged before we can harness manipulationism to mechanism.

Craver (2007) picks up this challenge, by arguing that, in mechanistic explanation, the hierarchical organization of mechanisms, which can also be modeled using invariant generalizations of a particular sort, are bearing some of the explanatory work. He offers an account of constitutive relevance that draws upon manipulationist concepts, which I turn now to consider.

## 3.3 Craver and Constitutive Explanation

Craver (2007) moves the rapprochement forward by drawing upon Salmon's (1984) distinction between etiological causal-mechanical explanation and constitutive causal-mechanical explanations. Where etiological explanation encompasses the prior causes of the explanandum, constitutive explanation encompasses the internal causes within the explanandum. Craver argues that while Woodward's account of manipulationism

[3]Woodward would, I think, allow that a causal structure over $A$ and $B$ is not a very deep explanation, and that uncovering a finer-grained causal structure would yield a deeper explanation. My point nevertheless stands.

provides a solid account of etiological explanation, mechanistic explanations do not generally look to the explanandum's etiology, but to their constitution. Where Woodward's **(MECH)** does not account for constitutive explanatory relevance, Craver's goal is to provide a characterization of constitutive relevance that can take advantage of manipulationist concepts.

Giving a constitutive explanation requires that we first identify the components of the mechanism—that is, those elements that are *constitutively relevant* to the explanandum—, and the causal organization of those components that make the explanandum phenomenon possible. Craver invokes, as did Woodward, manipulationism to describe the causal dependencies within a mechanism: Craver characterizes mechanism activities as Woodward-style generalizations invariant under interventions. A full mechanistic explanation will show how the joint action of the components bring about (in a non-causal sense that I explore below) the explanandum phenomenon, and how the invariant generalizations therefore can be harnessed for the purposes of prediction and control.

On Craver's view the components are not *causally* related to the explanandum phenomenon in any way; the proper relationship of component to explanandum is that of part to whole (see also Craver & Bechtel, 2007). So the joint action of the components *just is* the behavior of the explanandum phenomenon. Nevertheless, like the causal relation, the part-whole relation, Craver notes, also exhibits a pattern of counterfactual dependence whereby we can manipulate the whole by intervening into the parts, and *vice versa*. If I shorten each of the legs of a table, I have shortened the table; if I increase the sodium concentration within an neural axon, I have *ipso facto* altered the action potential's rate of propagation. This feature of constitution suggests a criterion for constitutive explanatory relevance, which Craver calls Mutual Manipulability.

## Mutual Manipulability

Whereas etiological explanation has a clear criterion for explanatory relevance—if $A$ is a cause of the explanandum $B$, then it is etiologically explanatorily relevant—there is as yet no clear criterion for constitutive explanation. Craver demonstrates that he can harness manipulationist principles as the basis for an account of constitutive explanatory relevance, which he calls Mutual Manipulability.

**(MM)** An entity $X$ engaged in activity $\phi$ is constitutively relevant to the behavior $\Psi$ of a mechanism $M$ [the explanandum phenomenon] if and only if

(i) $X$ is a (spatio-temporal) part of $M$, (Craver, 2007, p. 153)

(ii) when $\phi$ is set to the value $\phi_1$ in an ideal intervention, then $\Psi$ takes on the value $f(\phi_1)$ (p. 155), and

(iii) if $\Psi$ is set to the value $\Psi_1$ in an ideal intervention, then $\phi$ takes on the value $g(\Psi_1)$ (p. 159).[4]

Conditions (ii) and (iii) together capture the idea expressed above that parts and wholes enter into relations of manipulability. Given a whole, its total set of parts is identical with it. Because of this relation of identity, changes in the parts just are changes in the whole, and vice versa. Constitution is not a causal relationship, and yet it is a relationship that yields relations of manipulability (again, see also Craver & Bechtel, 2007, for a discussion of this distinction). This feature of constitution permits a test for constitutive relevance: If a putative component is mutually manipulable with the explanandum phenomenon, then it is constitutively relevant to the explanation for that phenomenon.

[4] Craver (2007) uses $S$ to represent the mechanism. I will use $M$ to represent the mechanism, as in subsequent chapters, I will require $S$ to represent the start or input conditions. For the sake of consistency, I have re-lettered Craver's usage to match my later usage.

Although **(MM)** condition (i) may appear superfluous, it is necessary for **(MM)** to work. It is necessary because an entity that causes the explanandum, and is also caused by the explanandum will exhibit mutual manipulability with the explanandum, and yet remain irrelevant to the explanation. Imagine a cognitive phenomenon as performing mental calculations. A subject's heart beat, then, will have a bearing on the subject's ability to calculate (if it is too fast, calculations will turn up incorrect more often than not, suppose; if it is too slow, the calculations will not be performed as the subject passes out from lack of oxygen). Too could the calculations have an effect on heart beat (if, for instance, the subject will be rewarded for correct calculations, or if the calculations are particularly vexing). Yet the heart rate is surely not relevant to a mechanistic explanation of this cognitive phenomenon: It is not relevant, because it is not part of the mechanism.

## Manipulability as a Sign of Causation and Constitution

Where Woodward's quantitative account of mechanism lost sight of the central features of the extant descriptive accounts, Craver's quantitative account of mechanism runs into problems with manipulationist formalism. I raise two issues for Craver. First, I worry that his account irretrievably conflates causation and constitution by treating both as relations of counterfactual dependence invariant under interventions. Second, I worry that, even if the first worry is fully addressed, the strong hierarchical and mereological features of his account defy tidy representation in a formal model, and hence defy analysis by the tools of causal modeling—which was one of the very goals of rapprochement.

Here is my first worry in detail. Woodward takes an observed relationship of manipulability as a sign of causation. Central to the manipulationist project is that if when wiggling $A$ we observe $B$ to wiggle, this observation is evidence that $A$ is

causally relevant to $B$. Central to Craver's project is that if when wiggling $A$ we observe $B$ to wiggle, this observation is (partial) evidence that $A$ is constitutively relevant to $B$. Suppose, then, that we wiggle $A$, and observe $B$ to wiggle. What should we conclude? Our options are: That $A$ is causally relevant to $B$, that $A$ is constitutively relevant to $B$, that $A$ is both causally and constitutively relevant to $B$.

Craver offers a principled way for making this decision. First, he categorically denies the possibility of inter-level causation (a point emphasized by Craver & Bechtel (2007)). Second, he observes that constitutive relevance is a symmetric relationship, where causal relevance is not: If $A$ is constitutively relevant to $B$, then if we were to wiggle $B$, we should expect to observe $A$ to wiggle. If $A$ caused $B$, we should not expect that.

Thus, constitutive relationships are always *mutually* manipulable, the relation of manipulability is symmetric. Causal relations, on the other hand, are asymmetric: Manipulating the cause will change the effect, but not *vice versa*. However, there is no reason to think that a mechanism could not comprise parts that are causes each other, where $A$ causes $B$, and $B$ causes $A$. Homeostatic mechanisms, mechanisms with feedback loops, are one example. A thermostat works by simultaneously measuring the air temperature, and adjusting a heating element to bring the the air temperature to some set value. The heating element and the temperature sensor exhibit, therefore, a relationship of mutual manipulability, yet neither is constituted by the other.

Craver offers a principled way to distinguish constitutive relevance from feedback loops in **(MM)** condition (i) above. Condition (i) stipulates that for $A$ to be constitutively relevant to $B$, $A$ must be a part of $B$ in a straightforward spatio-temporal sense. But this response only begs the question off a little longer, because this spatio-temporal notion of parthood already presumes some notion of system

boundaries—which is precisely the issue at stake.[5]

One possible way to assess parthood is by inspection: A car has wheels, carburetors, doors, and so forth. All of these are parts of the car. Naturally, not all of these parts are relevant to particular mechanistic explanations of the car, *e.g.* the mechanism for acceleration, but that's hardly a problem because once we have the candidate parts, mutual manipulability will take care of determining which are mechanistically relevant. Such a response, however, amounts to an appeal to compartmental boundaries or spatial coherence, principles Craver has already rejected (see his Chapter 5): How else do we know which parts belong to the car, except to observe that there is some set of parts that literally hang together in a particular way (coherence), and which are contained roughly within the metal or plastic panels encasing it (compartments)?

Thus, without a principle that can independently identify the spatio-temporal parts of the mechanism in hand, we cannot distinguish feedback loops from relations of mechanistic constitution—both are relations of manipulability. Even if we had such a principle, however, I have a second worry.

## Representing Interlevel Interventions

My second worry is that there is no coherent representation for Craver's constitutive conception of mechanisms using causal Bayes networks. Recall that one of the motivations for rapprochement was to harness the power of formal causal inference, which requires that we be able to create explicit representations of a mechanism within the formalism of causal graphs. But how do we represent constitutive relevance?

One option is with arrows. Because constitution and causation both exhibit patterns of counterfactual dependence invariant under intervention, and because Wood-

[5]With thanks to Craver for offering a tidy way to make this point.

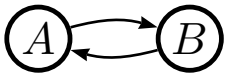


Figure 3.2: A causal model over $A$ and $B$ that is ambiguous with respect to whether $A$ is a part of $B$, $B$ is a part of $A$, or $A$ and $B$ are engaged in a causal feedback loop.

ward's semantics for causal models has arrows representing such relationships, both constitutive and causal relevance relations could be reasonably represented in a causal graph by arrows (even with a principle for determining parthood).

But using the same formal element to represent two very distinct ontological categories creates onerous difficulties, for our inferences will rely only on the formal properties of the graph; if the graph does not distinguish causal from constitutive relevance, then neither will our inferences. If we have a graph such as in Figure 3.2, we cannot tell from the graph alone whether $A$ is constitutively relevant to $B$, $B$ is constitutively relevant to $A$, or $A$ and $B$ form a causal feedback loop. So, on this representation, the previous worry continues to find purchase in the formal context. Perhaps there is another way to represent constitutive relevance.

On Craver's constitutive view, individual mechanisms span two levels: the level of the whole, and the level of its parts. Moreover, drawing on Glennan's hierarchical view of mechanisms, the parts may themselves be mechanisms, and the whole may be a part in a greater mechanism. Thus does Craver define his notion of *mechanistic levels*. Perhaps we should represent a mechanism with two levels, that of mechanism part and that of mechanism whole.

In fact, Craver does use this kind of representational schema, as exemplified in Figure 3.3. In this diagram, the mechanism $M$ sits over and above its components $X_1, \ldots, X_4$, representing that $M$ is at a higher mechanistic level than its parts. The arrows connecting the parts, $\phi_1, \ldots, \phi_4$ represent the activities that link the parts.

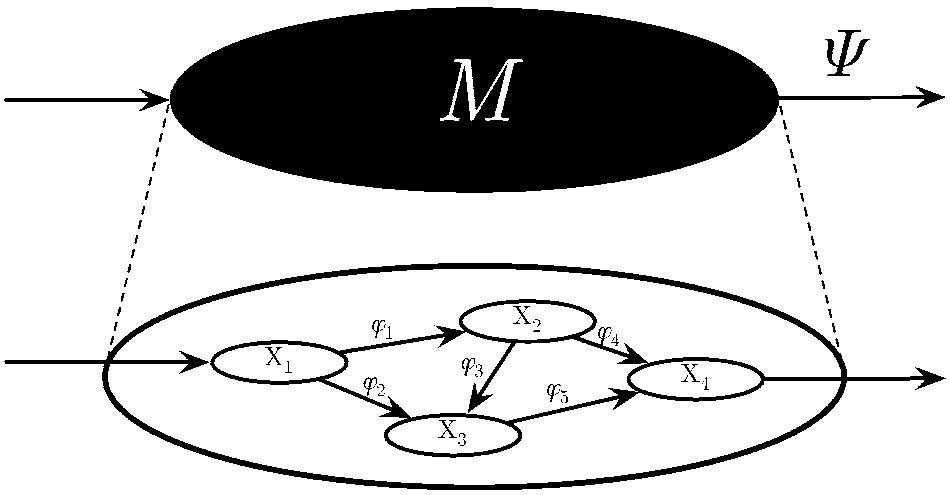


Figure 3.3: Schematic causal model of a mechanism. Nodes represent components; arrows represent activities. Adapted from Craver 2007, Figure 1.1, p. 7.

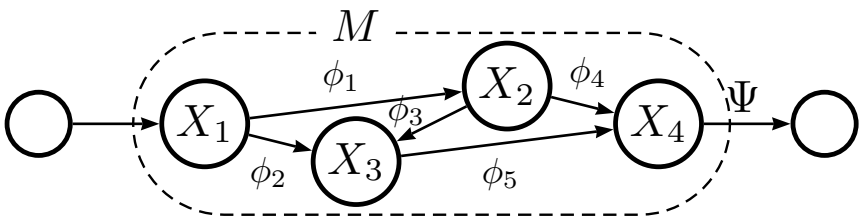


Figure 3.4: Redrawing Craver's figure, with identical elements collapsed for clarity.

But what are we to make of the double sets of arrows pointing into $M$ and $X_1$, and pointing out of $M$ and $X_4$? Do the pairs represent the same causal interaction, or distinct causal interactions? It cannot be the latter, because that would allow that we could change $M$ independently from $X_1$ in violation of **(MM)** conditions (ii) and (iii) above. So the pair of arrows into $M$ must both represent one and the same causal relation; and the same for the pair of arrows out. I make this identity explicit in Figure 3.4.

Presuming the representation in Figure 3.4 is suitable, how then are we to represent interventions into $M$? As I see it, there are (exactly) four ways to understand an

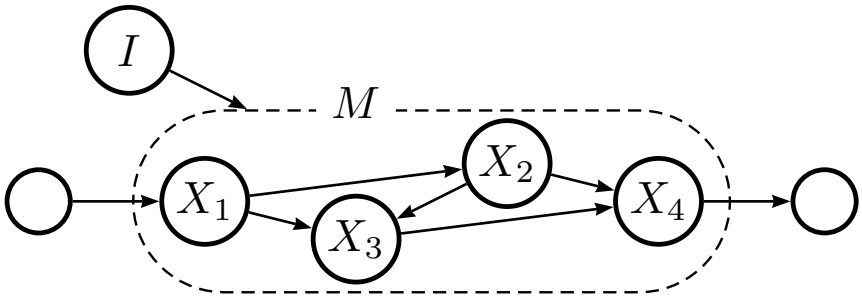


Figure 3.5: Intervening directly into $M$. (In this figure, and those following, I have for clarity left off the activity labels '$\Phi$', and '$\psi_1$'–'$\psi_5$'.)

intervention on $M$. These are presented diagrammatically in Figures 3.5–3.8.[6] Note that I do not want to beg any questions about how the effects of the parts bring about $M$'s $\psi$-ing, so there are no arrows connecting the mechanism parts to $M$ (*i.e.*, let us presume that we have found a solution to the first worries raised above about the conflation of causation and constitution.) I represent an intervention into $M$ as a node $I$, with an edge directed towards the intervened-into entity. Since intervention is a two-place predicate, suppose that these interventions are made with respect to $X_2$—that is, suppose that we want to know whether $X_2$ is a proper component of $M$.

In Figure 3.5, I construe 'intervening into $M$' literally. In fact, I am not sure how to intervene directly and literally into $M$ without requiring the $M$ have an independent existence from its parts, a thesis that denies that the whole is identical with the (organized) parts. How do I intervene into my lawnmower's mowing, or the neuron's firing if that intervention isn't in fact into one (or more) of my lawnmower's or neuron's parts? I set this issue aside, so as to avoid begging questions against Craver. Let us allow that it may well be possible to intervene directly into $M$.

If it is possible for to intervene directly into $M$ independently of intervening into its parts, then the intervention must be *causally independent* of all of the parts $X_i$: The

[6]The observant reader may wonder that there appear to be more than four ways, but these additional ways are degenerate cases that devolve into one of the four given, and so need not be considered additional cases.

intervention must be a cause of $M$, and must cut off all of $M$'s causes, including (if indeed they are causes of $M$) the $X_i$. The intervention cannot be a direct cause of $X_2$ (the dependent variable) on pain of conflation. But if the intervention is independent of the parts, and the behavior of the mechanism just is the organized interactions of its components, then it is hard to see how such an intervention would have an effect on $X_2$, or indeed, on any of the parts, unless there is a relationship of top-down causation from the mechanism whole to the individual parts.[7]

To use an example, Imagine a Rube Goldberg machine whose behavior is pouring my coffee. The coffee-pouring is an effect that occurs as a result of the various steps in the machine: the coffee-pouring is not contemporaneous with any subset of its parts' activities: neither the cat chasing the mouse into the hole, nor the mouse eating the cheese, nor the diminishing quantity of cheese tipping the scales. The tipping scale raises a lever that tips the coffee pot: only then does the coffee pour. An intervention into the coffee-pouring that is not simply an intervention into the machine's parts must occur at this point, cutting the rest of the machine off from the coffee-pouring.

The difficulty for imagining an intervention into $M$ in this way is that it requires $M$ to be a distinct entity from its parts. But Craver identifies the mechanism with its parts: Therefore an intervention into $M$ *must be* an intervention into one or more of its parts. So talk of direct interventions into $M$ is incoherent. Let us try a different way of representing an intervention into $M$, then.

So, perhaps an intervention into the mechanism is an intervention into one (or more) of its parts. But which part should we consider intervening into? Perhaps we should intervene into the part most causally downstream in the mechanism, the part that is the most direct cause of $M$'s $\Psi$-ing (in this case, $X_4$). Insofar as the activity of the whole is constituted by the activity of the parts, interventions into the behavior

[7]And, to reiterate, Craver is committed against the existence of such top-down causation (and again, see also Craver & Bechtel, 2007).

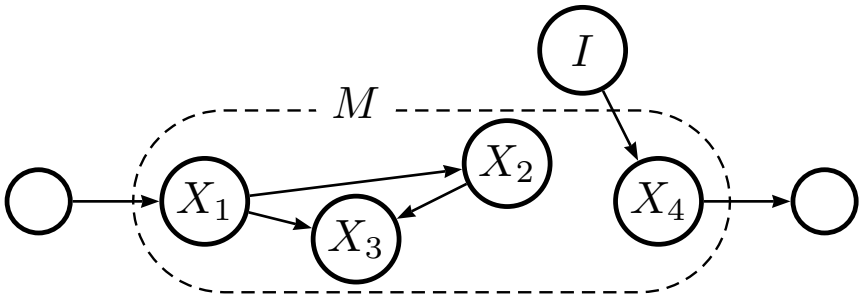


Figure 3.6: Intervening into $M$ via $X_3$.

of the whole should prefer interventions into parts closest to the production of that behavior, that is, those parts most causally downstream from the input conditions. Ideally, if we want to intervene into a lawnmower's mowing, we should consider intervening into the blades first, insofar as mowing is cutting grass, and the blades are what cut the grass directly.[8]

Suppose we intervene into the most causally downstream part, that closest (in some sense of the word) to the output of the mechanism. Figure 3.6 illustrates this case. But in this example, how could such an intervention affect any other component of the mechanism? Recall that we want to know if $X_2$ is a part of the mechanism; we are intervening into the mechanism to see if we can manipulate $X_2$, and so satisfy **(MM)** condition (iii) of **(MM)**. Because the intervention cuts all of the incoming arrows to $X_4$, and there are no edges from $X_4$ to $X_2$, there is no way for the intervention to manipulate $X_2$ (the dependent variable). So interventions into $M$ with respect to $X_2$ cannot proceed this way, as **(MM)** condition (iii) could never be satisfied: There is no causal route from $X_4$ to $X_2$.

Perhaps then, we can intervene into $M$ with respect to $X_2$ by intervening somewhere causally upstream of $X_2$. Figure 3.7 illustrates such an intervention, at $X_1$.

[8]In fact, it seems that we cannot intervene into any other part without foreknowledge of at least some of the mechanism. Knowing nothing of the mechanism of the lawnmower, we cannot even contemplate intervening into the fuel flow or carburetor. This is a serious issue, but a tangential issue; I set it aside.

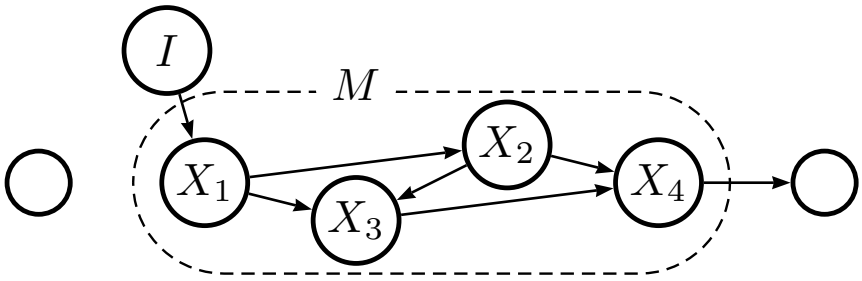


Figure 3.7: Intervening into $M$ via $X_1$.

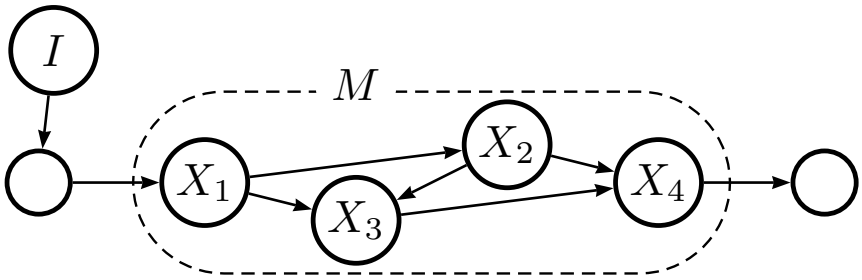


Figure 3.8: Intervening into $M$ via the start conditions.

In this case, condition (iii) will hold for $X_2$, but at dear cost, for the intervention is no longer ideal. The intervention manipulates $M$ through a route that includes $X_2$, which is strictly forbidden.[9] To see this, the intervention also causes $X_3$ and $X_4$, which partly constitutes (and hence changes) $M$. But the intervention is supposed to be into $M$. We have put the cart before the horse.

Finally, we might intervene into $M$ by intervening into the input or start conditions, as illustrated in Figure 3.8. But this kind of intervention fares no better than an intervention into $X_1$, and for the same reasons.

We have exhausted each of the possibilities for intervening into $M$, and found each one highly problematic. The problem lies in Craver's reliance on constitution: His representation of the constitutive relationship reveals that talk of intervening into a mechanism with respect to its parts is incoherent.

[9]Recall from §1.2 that a causal structure of the form $I \to Y \to X$ cannot count as an intervention into $Y$ with respect to $X$.

Note that this conclusion has profound repercussions for neuroimaging studies. Consider a stereotyped neuroimaging study. A neuroscientist wishes to investigate a behavior $\Psi$, which is explained by mechanism $M$. A subject is given some base task to perform, during which hemodynamic measurements are made. This base task is designed to specifically not to elicit behavior $\Psi$, so as to provide a contrast condition. The subject is then given a task that elicits behavior $\Psi$; hemodynamic measurements are made. These measurements are compared against the measurements from the base task; any differences are presumed to be a part of mechanism $M$.

This paradigm fits neatly with the interpretation in Figure 3.8: the intervention is causally *prior* to $M$: we might think of the task instructions as manipulating the neurons in the sensory system, which is taken to be prior to $M$ in some sense. This is taken, on Craver's model, to be a manipulation of $M$ with respect to $X$ (where $X$ also leads to changes in the brain's hemodynamics[10]. However, as pointed out above, this manipulation is not ideal, in that it manipulates $M$ on a route that passes through the component $X$. Thus, on Craver's account, we cannot conclude that $X$ is a part of $M$. We can only conclude that the task instructions and stimuli are a cause of $X$. For example, it might be the case that the task is a common cause of $X$ and $M$. Because the intervention (as Craver envisions it) fails to be ideal, we can conclude nothing about the relationship between $X$ and $M$.

The conclusion to draw from this discussion is that $M$ is not something above the parts, but a distinct entity, and requires a distinct representation in the graph. Condition (iii) of **(MM)** cannot be, therefore, tested—because $X$ is not a part of $M$. If the mechanism's behavior—the explanandum phenomenon—is something that is brought about by the activities and interactions of the parts of the mechanism, then the proper way to think about $M$ is as an effect of one or more of the components,

[10]for our purposes, we need not consider the possible confound involved in confusing $X$'s producing a hemodynamic change with $X$'s $\phi$-ing, which is ostensibly different.

that is, to identify it with what I have been calling the explanandum phenomenon, rather than the component parts.

Constitutive relevance, as I have shown, cannot be modeled in a causal graph. Insofar as Craver's account of mechanistic relevance as constitutive relevance cannot be modeled, his attempt at rapprochement does not quite get us a formal account of mechanism. In the next chapter, I consider a different kind of mechanistic relevance that does capture Craver's intuition that mechanism components are somehow parts of a whole, and show that this new relevance relation can be molded using causal graphs. The key is to discard the spatio-temporal parthood criterion in favor of a relation I call *causal betweenness*.

## 3.4 Conclusion

The time is ripe for a normative account of mechanistic explanation. We need a principled account of better and worse mechanistic explanation. Moreover, insofar as biologists are engaged in an epistemic enterprise, their norms of inference are fair game for evaluation; and hence too are the inferences themselves.

A good explanation includes all and only those elements that are explanatorily relevant; A good *mechanistic* explanation includes all and only those elements that are mechanistically relevant. Thus, a normative account of mechanistic explanation requires a principle to identify those elements that are mechanistically relevant. One way forward in this endeavor is to look to manipulationism.

Woodward (2002) took the first step forward with his manipulationist account of mechanism. Woodward identifies mechanistic relevance with his manipulationist version of causal relevance: An element is relevant to the explanandum if they exhibit a particular relationship of manipulability where intervening into that element

brings about changes in the explanandum. But, looking to the descriptive accounts of mechanism, this view appears to allow too much in: Not everything causally relevant to the production of the explanandum will be a genuine part of the mechanism for the explanandum, including background conditions and causal precursors to the mechanism.

Craver (2007) has taken a second step forward by offering a different account of mechanistic relevance. Observing that mechanistic explanation is a variety of what Salmon called constitutive explanation, Craver identifies mechanistic relevance with a manipulationist account of constitutive relevance: An element is relevant to the explanandum if they exhibit a relationship of mutual manipulability *and* the element is a spatio-temporal part of the explanandum. Craver's account of mechanistic relevance seems to get our intuitions about explanatory relevance right, but cannot be formalized using causal graphs. Because constitutive relevance cannot be adequately formalized, we cannot use the formal methods of manipulationism Craver appeals to to draw inferences about constitutive relationships.

The next step forward then is to develop an account of mechanistic relevance that does justice to the descriptive accounts of mechanism, and that can be formalized without remainder. In the next chapter, I will present and argue for an etiological conception of mechanistic relevance based on Woodward's causal relevance and a principle I call *mechanistic betweenness*.

# Chapter 4

# The Manipulated Mechanism: A New Rapprochement

I argued in the previous chapter that we need a normative account of mechanism. One way forward is to bring the descriptive accounts of mechanism into rapprochement with a formal framework for drawing inferences about mechanisms using causal models. In the previous chapter, I examined two steps towards rapprochement with manipulationism offered by Woodward (2002), and by Craver (2007). In this chapter, I present a step forward towards rapprochement, by presenting and arguing for two new elements: A semantics for mechanism models that I call the Interactionist View, and a principle of mechanistic relevance called Etiological Mutual Manipulability. I demonstrate that these elements create a rapprochement that is descriptively adequate, by using them to illustrate two real-world mechanisms: the mechanism for pencil sharpening and the mechanism for the neuron depolarization.

In this chapter, I present and argue for *the Manipulated Mechanism*, a rapprochement of mechanistic explanation with manipulationism built upon a principled semantics for mechanism models and a principle for identifying the explanatorily relevant components of a mechanism.

The key to successful rapprochement is a bridge between mechanism descriptions and mechanism models. This chapter is devoted to building just such a bridge. Two elements necessary for a bridge are missing in the current state of the rapprochement of mechanistic explanation and manipulationism: A semantics for mechanism models, and a mechanistic relevance principle that identifies which components/variables are properly considered relevant to a mechanism/mechanism model.

Glennan (2005) presents a set of desiderata meant to establish when a model of a mechanism model is adequately similar to the mechanism it represents. He asks:

1. Does the model predict (quantitatively or qualitatively) the overall behavior of the mechanism? Do these predictions hold for all inputs, or only for some ranges?
2. Has the model identified all of the components in the mechanism? Have the components been localized?
3. For each component, has the model correctly identified its causally relevant properties—that is, the properties whose changes figure into interactions with other components?
4. Does the model provide quantitatively accurate descriptions of the interactions and activities of each component?
5. Does the model correctly represent the spatial and temporal organization of the mechanism?

6. If the model includes sub-models of the mechanical structure of components, are these sub-models good representations of these components?

7. Is the mechanism identified by the model the sole mechanism responsible for the production of the behavior, or are there multiple mechanisms? If there are multiple mechanisms, do they operate concurrently and redundantly, or do different mechanisms operate in different contexts? (Glennan, 2005, p.457)

Ignoring the first question, which is a question that must be put to *any* causal model, Glennan's thought is that a candidate mechanism model must have answers to the remainder to qualify *as* a mechanism model. A mechanism model must, in other words, adequately capture certain central features of a genuine mechanistic explanation. It must account for all and only the component parts of the mechanism, the activities or interactions that bind them together as a mechanism, and the structure of the mechanism generally. (Already one can see that Woodward's (2002) account does not have the resources to answer these kind of questions.)

But what we do not yet have is a principled way of evaluating a model to see how well it can supply answers to these questions. Under what conditions can we say that a model has identified a particular component, or its causal relations? How do the formal elements of causal models map onto the qualitative elements of mechanistic explanations?

In this chapter, I develop principles for addressing Glennan's desiderata. In §4.1, I claim that mechanisms place constraints on the semantics for mechanism models. Mechanisms are entities and activities or interactions, productive of regular changes, organized in a particular way. If a causal model is to represent a mechanism, there

must be some interpretation of that model such that we can reconstruct a story about the entities and their activities and their organization from that model. This requirement places constraints on the possible semantics for the nodes and the arrows. What I take to be the Default View is that the nodes represent entities, and the arrows activities, but I will argue that this view is mistaken, because it leaves off the idea that the entities and activities *together* are productive of regular changes—and productivity is here doing the causal work. Thus, I will argue for a different view, the Interactivity View, that nodes represent what I call activity-component pairs, and arrows relations of manipulability.

There is more. As I will discuss in §4.2, mechanisms place constraints on the structure of causal models. I argue that a mechanism is a kind of causal explanation that links an explanandum cause to an explanandum effect—and that even apparently non-etiological mechanisms (constitutive mechanisms, homeostatic mechanisms, correlative mechanisms) can be redescribed in purely etiological terms. I call such a mechanism that links a cause to an effect an etiological mechanism. An etiological mechanistic explanation analyzes the explanandum causal relation into many lower-level[1] components that engage in lower-level causal relations. Thus, mechanisms are not *just* causal structures, but causal structures with a specific kind of organization, one that shows how an explanandum cause causally contributes to an explanandum effect. From these constraints on the causal structure of mechanisms, I derive a mechanistic relevance principle called Etiological Mutual Manipulability: The parts of a mechanism are all and only those components that satisfy the constraints on the causal structure of mechanisms.

Finally, in §4.3, I argue that a rapprochement using these semantics and mechanistic relevance principle can maintain the descriptive adequacy of the existing de-

[1]I use this word very loosely here. See Craver (2007, Chapter 5) for a detailed discussion of mechanistic hierarchies.

scriptive accounts of mechanistic explanation by demonstrating that they capture the relevant descriptive details of two mechanistic explanations, the mechanism for pencil sharpening and the mechanism for neuron depolarization during an action potential.

## 4.1 Semantics for Mechanism Models

Why do we want a semantics for mechanism models? A rapprochement is a bridge between descriptive and formal—it must show how the formal elements of causal modeling map onto the descriptive elements of mechanistic explanation. A rapprochement that fails to provide principles for inter-translation is not a genuine rapprochement.

Mechanisms comprise components, and these components engage in a variety of activities (in the Machamer, Darden, & Craver (2000) sense of the word). Constructing a rapprochement that maintains the descriptive adequacy of mechanistic explanation requires a carefully articulated link between the formal elements of a causal graph and these qualitative elements. This link is a semantics for causal graphs; specifically, it is a principle for the interpretation of arrows, and a principle for the interpretation of nodes.

Of course, a causal graph is a graph that has already been given some minimal interpretation of the nodes and arrows. The nodes represent causal relata, and the arrows represent causal relations. But accounts of mechanistic explanation have their own additional ontological commitments about what the causal relata can be, and manipulationism presents its own constraints on the causal relations, and so we need additional constraints on the semantics. We might take the nodes as representing mechanism components, and arrows as representing the causal interactions among the components—a position I call *the Default View*, because it is quite natural, and appears fairly frequently, if implicitly, in the mechanistic literature. I am going to

argue, however, that we should instead take a view I call *the Interactivity View*, in which we take the nodes as representing activity-component pairs—I call these 'active components'—, and the arrows as the causal interactions among the active components.

Before I turn to consider these two views, let us set the stage by considering some central constraints on a semantics for mechanism models.

## Components and Activities

There are two clear constraints on the possible interpretations of a graphical model of a mechanism. First, such models must somehow capture the idea that mechanisms are decomposable into component parts. Second, they must somehow capture the idea that the components are active and interact, and that these activities or interactions are the producers of change.

Common to all qualitative accounts of mechanism is that mechanisms are composed of *parts* or (my preferred term) *components*.[2] In most concrete examples of mechanisms, components are *entities* or *things*: middle-sized physical objects that occupy a finite but non-zero volume of space-time[3], *e.g.* levers, pumps, and volumes of water (Glennan, 2002); inclined planes and blocks (Woodward, 2002); mRNA and amino acids (Darden & Craver, 2006); ion channels (Craver, 2007). Sometimes, authors will admit more esoteric entities, *e.g.* electro-magnetic fields (Glennan, 2002) or voltage gradients (Craver, 2007) (neither of which occupies space in the sense of excluding other entities from occupying the same location). Glennan (2002) is more

[2]A terminological note: I use 'components' as a generic term for the kinds of causal relata found in a mechanism. I use 'part' in the mereological sense, *e.g.* 'part of a mechanism'.

[3]That said, most authors will grant fundamental particles—which are not middle-sized, and may not occupy non-zero volumes of space-time—as components. Eberhardt (personal communication) has pointed out that economic phenomenon do not seem located in space, in which case it looks like these accounts preclude the possibility of economic mechanisms.

explicitly ecumenical, in allowing that parts need not be localizable or describable in a purely physical vocabulary, and suggests software components and information could also be parts of mechanisms.

The idea that a mechanism can be decomposed into components is central to the descriptive accounts of mechanism; the behavior of a mechanism is generated by and only by the components of that mechanism (see Chapter 2). Thus, a large part of mechanistic explanation consists in identifying which of those components are genuine parts of the mechanism. Identifying a mechanism's components is the problem of *mechanistic relevance*. [4]

On the view of Machamer, Darden, & Craver (2000), the constituent components are *active*.[5] Machamer, Darden, & Craver make the point more forcefully when they claim that "mechanisms do things. They are active and so ought to be described in terms of the activities of their entities" (p. 5). Moreover, "activities are the producers of change" (p. 3). That is, the activities are the causal links between the components; without activities there would be no change, and without change there would be nothing to explain (because there would be no explanandum behavior). Activities, on this dualist view, are real things, as real as the components that engage in them.[6] More recently, Bogen (2004, 2008) and Machamer (2004) have offered independent defense of this dualist view.

[4]Of course some, perhaps Sandra Mitchell, might be quick to point out that biologists often talk about the 'mechanism for natural selection', which seems to have no particular components, not in this sense anyway: It's not like there is some enduring machine that sits idle, roaring to life only to (naturally) select one or more organisms. Whether natural selection is a genuine counter-example, or is simply not a mechanism, is open to debate. I do not mean to beg any questions, but to simply take note of a prominent feature of extant accounts of mechanism; whether and when it is decided that this feature requires modification, my account should naturally follow suit.

[5]Although Craver has since disavowed activities, being rather vexed entities

[6]I ignore that sometimes a lack of change, or a lack of behavior, is what needs explaining, *e.g.* homeostatic systems. Why, for example, is it that human bodies typically do not deviate very far from a fixed internal temperature? But homeostatic mechanisms are indeed composed of active components; without that activity, the static state *would* change; our bodies would cool off (or heat up) to match the ambient temperature. So even in these cases, activities are crucial to the explanation.

Where MDC maintained an explicit dualism, holding that activities had an existence independent of components, other authors have found this dualism at best vexed (*e.g.* Glennan, 2002; Craver, 2007). Nevertheless, it is a central feature of all accounts of mechanism that the components are not simply inert, but act together in a coordinated, productive way. Bechtel & Abrahamsen (2005) require that mechanisms contain "component operations" (p. 3), which they characterize as similar to activities, but with a greater emphasis on the idea that operations belong to individual components, and have no independent existence (a point Machamer, Darden, & Craver do not deny). Glennan (1996, 2002), who explicitly denounces activities, nevertheless claims that a mechanism "produces [the explanandum] behavior by the *interaction* of a number of parts" (p. S344, emphasis mine). Woodward (2002) and Craver (2007) likewise reject activities in favor of talk of invariant generalizations. Where interactions or invariant generalizations are expressed as purely quantitative relationships, on the manipulability view an activity is not an independent thing in its own right, but a description of the quantities and their functional relations.

Even if activities are falling out of favor as an analysis of causation, our models could only benefit from capturing the descriptive depth of activities. And if the dualist views of MDC, Bogen, and Machamer are right, then our models would be incomplete without a representation of activities. Thus, mechanism models shoud capture not just that this gear's position is a cause of that gear's position, but that this gear's *rotating* (an activity of the gear) is a cause of that gear's rotating. If we accept activities as ontologically *sui generis* and an indespensable part of mechanistic explanation, there must be room for them in the model. If we do not accept activities as anything but a kind of narrative description of unanalyzed causal relations, our models would still be improved if they could include some kind of referent to these descriptions. So, I include activities among the items a mechanistic semantics must capture, without

committing myself to the dualist view.

Let us see now one common semantics for models, and how it fares against these considerations.

## The Default View

On one common semantics for graphical mechanism models, the arrows represent activities (the producers of change), and variables represent mechanism components. I take this to be the Default View, because it is quite natural, captures the descriptive constraints described above, and in the absence of any explicit interpretive strategy, this is the view that most authors seem to reflexively adopt.

**Default View** Variables represent components and arrows represent activities or interactions.

The Default View has the virtue of being very natural, in that it brings out and exploits an obvious structural similarity between causal models and mechanisms: Mechanisms have components and activities, graphs have nodes and arrows. This view is so natural that several authors (including Machamer, Darden, & Craver, 2000; Woodward, 2002; Craver, 2007) have taken it independently, if reflexively. Machamer, Darden, & Craver write that "If a mechanism is represented schematically by $A \rightarrow B \rightarrow C$, then the continuity lies in the arrows and their explication is in terms of the activities that the arrows represent" (p. 3).

Woodward (2002) says that a mechanism model must "(i) describe an organized or structured set of parts or components, where (ii) the behavior of each component is described by a generalization that is invariant under intervention" (p. S375). Again, since manipulationism has it that arrows represent relationships of manipulability, Woodward here seems committed to the additional claim that arrows represent the

behaviors of components—where I read 'behavior' as closely analogous, if not synonymous, to 'activity'. Less directly, he is committed to taking nodes as representing components, in that he requires components have a representation in the model (by (i)), and that, with arrows having received an interpretation, only nodes remain to assign meaning to.

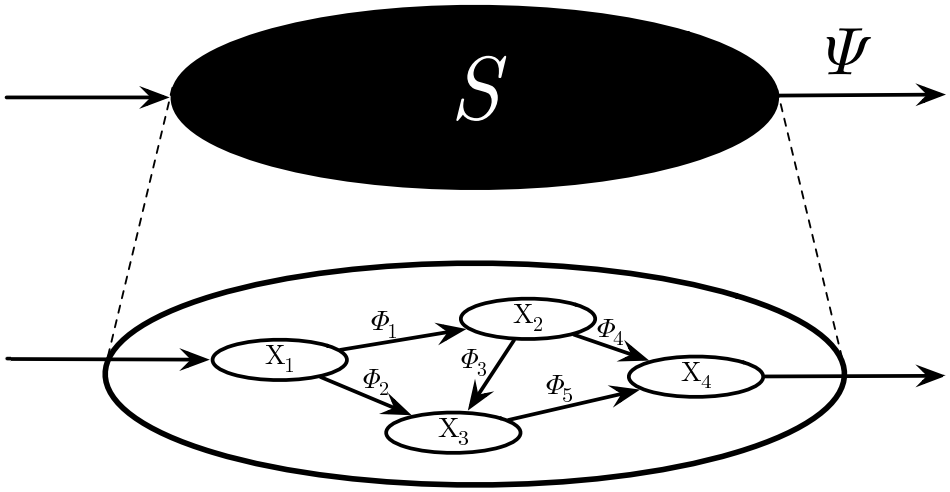


Figure 4.1: Schematic causal model of a mechanism. Nodes represent components; arrows represent activities. Taken from Craver (2007, Figure 1.1, p. 7).

Craver (2007) frequently (though not consistently) constructs mechanism models using the Default View (see Figure 4.1). Craver describes this figure:

> At the top is the phenomenon to be explained.... Beneath $S$'s $\psi$-ing are represented the entities [components] (circles) and activities (arrows) that are are organized together in the mechanism. (p. 7)

Here is a simple example applying the Default View. Suppose that we wish to model the mechanism for protein synthesis. Darden & Craver (2006) describe (very schematically) the mechanism in this way:

> The base sequence in DNA is transcribed into messenger RNA, which moves to the ribosomes, the site for the subsequent stages. A specific triplet codon on the messenger RNA hydrogen bonds to its complementary anticodon on a transfer RNA, which is attached to its specific activated amino acid. As the transfer RNAs bond sequentially to the messenger RNA, the amino acids are brought into appropriate proximity so that peptide bond form. Incorporation of amino acids occurs in a specific linear order, based on the order of the codons in the messenger RNA. (p. 82)

The components and activities are easy enough to see. Minimally, a model of this mechanism should include mRNA, tRNA, and amino acids. Activities include mRNA transcoding onto tRNA, tRNA hydrogen bonding with amino acids, and amino acids peptide bonding to each other. On the Default View, then, a graphical model of the mechanism for protein synthesis might look like Figure 4.2.

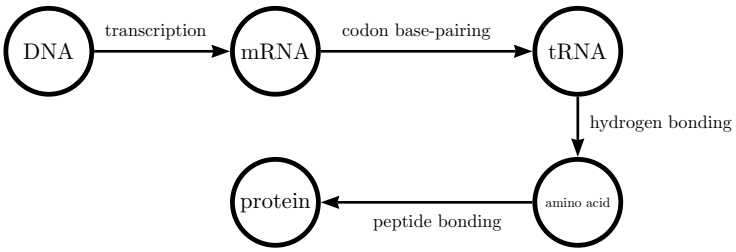


Figure 4.2: Schematic model of the mechanism for protein synthesis, with Default View semantics.

In this figure, I have chosen the nodes to represent the four components identified above, and the arrows to represent the productive activities of the components. Indeed, as one might hope, this figure is not too different from elementary textbook accounts, or even Crick's (1970) central dogma (Figure 4.3). The similarities to actual

scientific texts suggests that perhaps this is the right way forward.

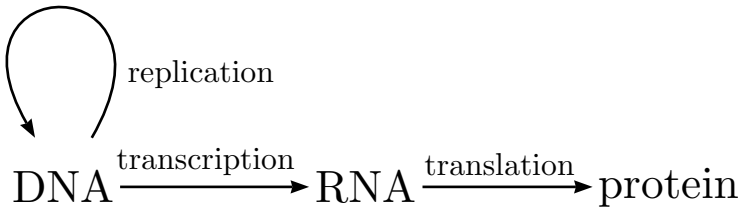


Figure 4.3: Crick's central dogma (adapted from Crick, 1970).

## A Problem with the Default View

But as natural and inoffensive as the Default View is, it cannot work, because it does not take the formal element of the rapprochement seriously enough. Montaña (2009) presents a compelling argument that the Default View is the wrong way to think about modeling mechanisms. Recall that graphical models capture the structure of the causal relations within a mechanism (or causal system generally), but not the fine details about the mathematical relationships among the nodes. The Default View creates an undesirable divide between graphical models and their functional analogs (*e.g.* systems of causally interpreted equations). Montaña asks us to take a graphical model such as the one in Figure 4.2, and use it to construct a functional model. He claims that we cannot, because components are not the kind of thing that can be assigned values, and hence be represented by the variables in a mathematical function. And if we cannot make this transformation, then something has gone wrong in our representation. He argues that what has gone wrong is the Default View's insistence on components as the causal relata, and thus urges us to rethink how to interpret mechanism models.

Here is Montaña's argument. He begins by noting that components (he calls them

'parts') and properties or attributes belong to distinct ontological categories. If we remove a component from a mechanism, we destroy it.[7] But attributes cannot be removed from a mechanism, at least, not without removing a component. We can change the color, for example, of one or more components, but we cannot *remove* color—color is not the kind of thing that can be removed. Moreover, Montaña notes, whereas attributes—being determinables—are the kind of thing that can take a value (*e.g.*, color can take a value from the range {violet, blue, green,...}), components are not the kind of thing that can take a value. What is the value of a wheel? Montaña thinks this question nonsense. We can ask the value of a wheel's position or its rotational velocity or its weight or its color, but not the value of a wheel full-stop. Thus, components and attributes are distinct ontological categories.

Next, Montaña observes that causal models are meant to capture the relationship between the values of its variables. Underwriting a graphical model is either a probabilistic or algebraic model, which relates the values of the variables. Therefore, in causal models, variables can only represent something that can take values, namely attributes. Since attributes are ontologically distinct entities from components, variables in a causal model cannot, therefore, represent components.

By extension, arrows in a causal model cannot represent activities. Recall that activities are always activities *of* components, there are no activities *simpliciter*. Thus, if variables cannot represent components, then it is not clear what the arrows should or can represent, except that it must not be activities.

One way to read Montaña's conclusion is that components just aren't the causal relata in a mechanism. Interestingly, despite their apparent commitment to the De-

[7]This is not to say that we destroy it utterly, but insofar as a mechanism is always a mechanism *for* a behavior, and removing a component leads to a cessation of that behavior, then what remains after removing a component is not that mechanism. It may well be that a new behavior emerges from the removal (imagine removing the wheels from a car; the car no longer moves, but it can still be used to *e.g.* generate electricity), but the original mechanism is gone.

fault View, Machamer, Darden, & Craver acknowledge this same point by observing that "...objects *simpliciter*...may be said to be causes only in a derivative sense" (p. 6) The reason is worth noting. "An entity," MDC tell us, "acts as a cause when it engages in a productive activity" (p. 6). Noticing that components are not causal relata (at least not as far as mechanisms are concerned), and noticing too that the variables of a causal model are the representatives of causal relata, why then should we accept the Default View? Clearly we shouldn't.

I turn now to address the shortcomings presented above in my own semantics for mechanism models.

## The Interactivity View: Activity-Component Pairs and Relations of Manipulability

Recall that Glennan (2002) characterizes mechanisms as the "interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (p. S344). Interactions are deflationary: To say that one component interacts with another is simply to say that the one *brings about a change* in the other. We are left to wonder: Yes, but how is that change brought about?

Machamer, Darden, & Craver characterize mechanisms as "are entities and activities organized such that they are productive of regular changes" (p.3) Activities are meant to supply an answer to this worry: To say that a component engages in an activity is to say that that an active component is productively engaged in a process of change. Activities are meant to be descriptions of what it means to 'bring about' in a particular context. But we are left to wonder: What are activities changing or bringing about?

Tabery (2004) asks both of these questions, and where Glennan and MDC took their views on interactions and activities to be mutually exclusive, Tabery cogently notes that they are anything but. They are complementary. Activities open up the interaction 'black box', and tell us *how* components interact. But activities need interactions too: "We must identify what... makes the producer productive" (Tabery, 2004, p. 11). It's fine to say that a component *e.g.* pushes, but without a clear notion of what this push is changing or to what degree, the activity of pushing is not informative enough. Filling those blanks requires an appeal to interactions.

Tabery's observations point the way forward in replacing the Default View. He concludes with a call to explore the exact relationship between interactions and activities. An account of this relationship will also, as I will show, provide a suitable response to the Default View.

"An entity," Machamer, Darden, & Craver tell us, "acts as a cause *when it engages in a productive activity*" (p. 6, emphasis mine). This is an interesting and different claim: The claim here is that the causal relata are not components (as with the Default View), but *activity-component pairs*. But if activities are part of the *relata*, what is doing the *relating*?

Glennan offers an option. On his account, interactions among components "can be characterized by direct, invariant, change-relating generalizations" (Glennan, 2002, p.S344). And recall again that these kind of generalizations reflect relationships of manipulability. On Woodward's (2003) manipulationist account, relations of manipulability are to be represented by arrows in a causal graph. Thus, to say that $X$ interacts with $Y$ is to say that we could bring about changes in $Y$ by manipulating $X$, or graphically, $X \rightarrow Y$. So, taking Glennan's view of the causal relation in a mechanism, *contra* the Default View, arrows represent not activities, but interactions.

If Tabery is right, and interactions and activities are complementary, then the

foregoing discussion suggests that the relationship between them must be this: Activities (with their bearers, components), being determinables, are the causal relata, and interactions are the causal relation. Interactions support counterfactual claims about the productive relations among activities. Thus, we should use variables to represent activity-component pairs (since activities are always activities *of* something), and arrows to represent the interactions among activity-component pairs.[8]

**Interactivity View** Variables represent activity-component pairs and arrows represent relationships of manipulability (counterfactual dependence under intervention).

One might worry that the Interactivity View runs into its own problem with the formal representation. As I discussed in Chapter 2, Psillos (2004) has argued that activities require counterfactuals. And I have just argued that activities should be represented by variables. But this seems a category mistake: Counterfactuals have modal content, where variables do not. There is nothing modal in the claim '$X = 5$'. I respond by observing that Tabery's synthesis of activities and interactions places the modal onus on interactions. The 'change-relating' aspects of MDC's concept of activities are better handled by the concept of an interaction, and the description of *how* the change is brought about is better handled by the concept of activities. In this way, it seems that thinking of activities as a quantifiable attribute of a component is the right way forward: The gear is *rotating* at such-and-such a rate; the enzyme is *catalyzing* at this-or-that rate. Described in this way, there is no dependence relation,

[8]Note that this view permits a component to appear in multiple places in a causal graph; components often have more than one activity, and several of these activities may be components in a mechanism, and hence need distinct representation. For example, in a given mechanism, a gear's rotation and its momentum (resistance to movement) may play two distinct roles in a mechanism, in which case we will need two variables in our model, both of which represent a distinct aspect of the same component, the gear. With thanks to Jan Plate and Dennis Des Chene for helping me see this point.

let alone a relation of counterfactual dependence, because there is only the actor, and not the subject. The counterfactual dependence enters only when we assign a subject to the verbs—when we view the activity as interacting with another component.

The Interactivity View maintains the proper links to the descriptive elements, without running into difficulties with the representation. Moreover, by explicitly linking relations of manipulability to a feature of the mechanism—interactions—we have explicitly incorporated not just causal modeling, but specifically manipulationism as a feature of the rapprochement.

However, the full extent of the usefulness of the Interactivity View won't become clear until we consider the structural constraints on mechanism models. I turn now to give these constraints, and then to show how these constraints when conjoined with the Interactivity view yield a complete formal account of mechanistic explanation.

## 4.2 Mechanistic Relevance

Not just any component or activity or interaction is relevant to a particular mechanistic explanation. My consumption of table salt last night, although a source of sodium, does not figure into an explanation for the movements of sodium across cell membranes in my nervous system, even as the consumption of table salt is perhaps causally relevant. In the previous chapter, I argued for the need for an account of mechanistic relevance, and against Craver's (2007) account of constitutive relevance; I turn now to consider a etiological approach to solving this problem.

I begin by observing a close tie between mechanistic relevance and the causal structure of mechanisms. Mechanisms have specific constraints on their causal structure. In particular, since mechanisms are posited to explain cause-and-effect relationships, the mechanism must have a causal structure that links the explanandum cause to the

explanandum effect *as* cause to effect. Any active component that is not part of the link from explanandum cause to explanandum effect is not going to be relevant to explaining how it is that the cause brought the effect about. On the other hand, any active component that *is* part of this link *will* be relevant. I consider these claims in more detail below.

On the other hand, causal graphs are quite flexible in how they can be structured. As a result, not just any causal graph can be used to represent a mechanism. In this section, I lay out the structural constraints imposed on graphical models by the descriptive accounts of mechanism. I will develop a principle, called M-Separation, that distinguishes causal structures that are genuine representations of a mechanism from causal structures that cannot be used to represent a mechanism. I will then argue that this principle can be used as the foundation for an etiological account of mechanistic relevance that I call Etiological Mutual Manipulability.

I begin by examining the structural constraints imposed by mechanistic explanation.

## Mechanisms link Cause to Effect

The problem of explanatory relevance is the problem of identifying just those components that are necessary for an explanation, and ignoring the components that are not. When we aim to explain the dissolution of some salt in water, we should appeal to certain aspects of the salt's crystalline structure, and the bi-polar nature of water molecules; we should not appeal to the salt's having been hexed, or being located in the northern hemisphere, or the dissolution's taking place at precisely 7:35PM. These later factors have no bearing on how or why the salt dissolves, because they are not causally related to the salt's dissolution. In the previous chapter, I introduced Salmon's causal-mechanical concept of explanatory relevance, which claims that only

causes explain their effects; neither do effects explain their causes nor do mere correlates explain each other.

Here is a claim that I want to defend: Mechanisms explain by linking explanandum cause to explanandum effect *as* cause to effect. Mechanisms are posited to explain cause-and-effect relationships; I call the cause in the causal relationship that needs explaining the explanandum cause, and the effect the explanandum effect. There are many ways that the explanandum cause could be conceivably linked to the explanandum effect: They might coincidentally co-vary; they might share a common cause; they might occur in some temporal sequence; they might both be the same color; one might be a cause of the other. If we want to understand why it is that the explanandum cause brings the explanandum effect about, then only explanations that analyze this causal link are going to be useful; explanations that appeal to a common cause, or a shared color are not going to be useful. We can take advantage of the claim that mechanisms link cause to effect to identify which components, among the many available, belong to the mechanism, and hence are relevant to a mechanistic explanation.

One might object to this characterization of mechanistic relevance, because some mechanisms (one might claim) do not exhibit this kind of etiological cause-to-effect structure. Craver (2007) distinguishes etiological mechanisms (mechanisms that do link an explanandum cause to an explanandum effect) from constitutive mechanisms[9], which explain by decomposition: showing how a complex phenomenon is constituted by—not caused by—the mechanism's parts. Glennan (personal communication) has made the further claim that homeostatic mechanisms—mechanisms that use feedback to maintain a steady state–and correlative mechanisms—mechanisms that explain two

[9]In the previous chapter, I considered Craver's principle for constitutive relevance, Mutual Manipulability, and ultimately rejected it; here, I am considering something slightly different, Craver's constitutive mechanisms, for which Mutual Manipulability was a relevance measure.

distinct effects in virtue of being their common cause—do not offer explanations of a cause-and-effect relationship, and cannot therefore be cast into this etiological mold. I will argue that all three of these mechanism types can be described equally well as etiological mechanisms, and hence cast as explaining the link between a cause and an effect. Once I have established this, I will lay out a principle of mechanistic relevance.

### Constitutive Mechanisms

A constitutive mechanistic explanation is an explanation of a complex phenomenon that shows the phenomenon is *constituted* by the parts of a mechanism. On this view, the mechanism is identical with the explanandum phenomenon (and hence that the components of the mechanism are proper parts of the explanandum), and the explanation proceeds by analyzing the phenomenon into interacting parts. On this view, there is not a clear cause and effect relationship that is being explained by the mechanism, but rather the phenomenon is shown to constitute some set of causal interactions. Usually, constitutive explanations are offered to explain complex phenomena, phenomena that exhibit a number of interesting features. The action potential in the neuron, to take one of Craver's (2007) example, has many facets: A particular time-dependent change in membrane potential; flows of sodium and potassium ionic currents; the propagation of the signal from one axon segment to another; and so forth. Together, these hang together as a single, coherent, but multi-faceted phenomenon. It is a single phenomenon, because there is a single mechanism responsible for it, that can explain all of its various facets.

In the previous chapter, I argued that one cannot represent an ideal intervention into a constitutive mechanism, that is, we cannot model what Craver calls a top-down intervention. The problem is that there is no way to intervene into a mechanism that isn't somehow an intervention into one or more of its components. I would like to

suggest another way to think about top-down interventions, from an etiological point of view. Craver offers an example of a top-down experiment. "In activation experiments," Craver writes (p. 151), "one intervenes to activate, trigger, or augment the explanandum phenomenon and the detects the properties or activities of one or more putative components of its mechanism". Neural imagine experiments are a kind of activation experiment: A subject is given a stimulus that activates some cognitive mechanism, *e.g.* a text passage or an image, and a scanner is used to detect the activity of neurons or neural regions in the brain. But notice that the stimulus is, in this case, clearly a cause of both the cognitive mechanism's activation, and of the neural activity—precisely because the cognitive activity is constituted by—*just is*—the neural activity.[10] Thus, the experimental intervention—the presentation of a stimulus—is causally prior to both the mechanism and the explanandum phenomenon.

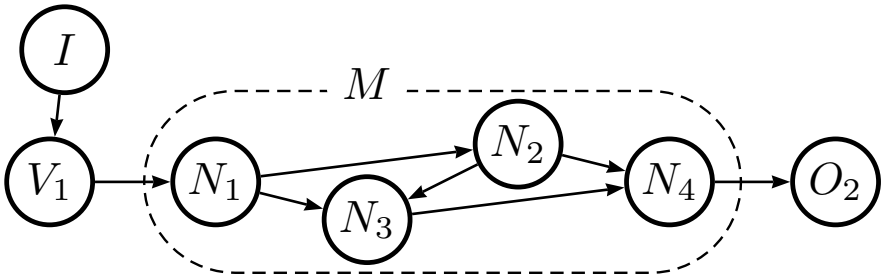


Figure 4.4: $V_1$ is the primary visual cortex; $N_1$–$N_4$ are brain regions that comprise a cognitive mechanism of interest; $O_2$ is the oxygenation of the blood in those regions.

Figure 4.4 illustrates one way we might model such an experiment. Suppose that the cognitive mechanism under investigation comprises brain regions $N_1$–$N_4$. We present a visual stimulus, activating the primary visual cortex $V_1$, which, through a number of processing steps in various explanatorily irrelevant brain regions, results

[10]On a physicalist view, anyway. I don't mean to beg questions against dualists.

in signals being passed into the mechanism. As these regions $N_1$–$N_4$ are activated, they (presumably) trigger an increase in blood oxygenation in the same physical location, $O_2$.[11] Our fMRI scanner detects this blood oxygenation.

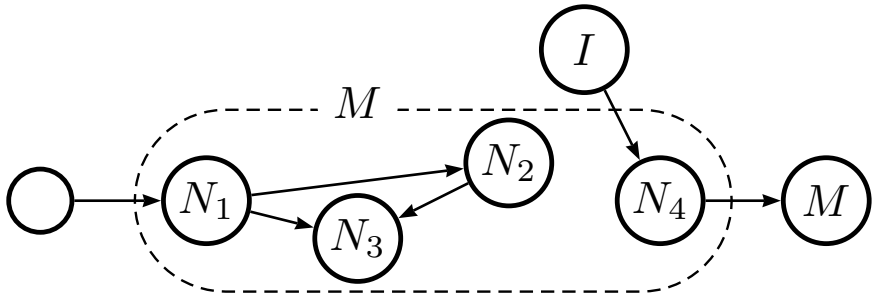


Figure 4.5: $N_1$–$N_4$ are brain regions in the motor cortex; $M$ is the movement of a particular muscle.

Notice that we can also easily give bottom-up experiments the same kind of treatment. In stimulation experiments, for example, "one intervenes to excite or intensify some component in a mechanism and then detects the effects of that intervention on the explanandum phenomenon" (p. 149). Fritsch and Hertzog's famous experiments introduced small electrical currents into the motor cortex to produce movements in specific muscles. Such an experiment might be modeled as in Figure 4.5. They intervened directly into a brain region using an electric probe[12]; the neurons stimulated sent a signal to a particular muscle (mediated, no doubt, by the cerebellum), causing the muscle to contract. This kind of experiment is easy to redescribe in purely etiological terms.

Notice that this description of an activation experiment does not make appeals to

[11] Really, there should be an $O_2$ node for each brain region $N_n$, but I wanted to keep the figure simple; my points will stand in either case.

[12] Whether their interventions were 'hard', or arrow-breaking interventions, or even ideal, makes this example just a little thorny, but let us suppose for now that these issues do not arise. I will deal with this kind of issue in Chapter 8

levels. This is not to say that Craver's hierarchical interpretation of these experiments is mistaken, only that there are multiple descriptions of these experiments, some of which are etiological, and some of which are inter-level.

Thus, we can re-describe a constitutive mechanism that explains some complex phenomenon as a complex of overlapping etiological mechanisms, each explaining a simple facet of the complex phenomenon.

### Homeostatic Mechanisms

Homeostatic mechanisms are mechanisms that use feedback to correct deviations from a particular state, thus maintaining that state. Homeostatic mechanisms are a bit different, because although they do have an etiological structure; what they lack are clear start and stop conditions, or explanandum causes.[13] We *could* ask, how was it on this occasion that the temperature dipped below the thermostat setting that the thermostat brought the temperature back up? Such a request specifies a cause (the deviant state), and the effect (the final state), and the explanation of this change will certainly be an etiological mechanism, but a *token* mechanism. Mechanisms are generally thought of as types, explaining a broad range of counterfactuals. And insofar as a homeostatic mechanism is a type, it appears to lack clear start conditions.

But we must be careful here not to conflate mechanisms with *machines*. A machine, as Glennan uses the term, is a collection of physico-mechanical parts that are connected together, but which are not themselves explanatory. A pencil sharpener, or an automobile are examples of machines. A machine becomes a mechanism only when it is part of an explanation for a phenomenon. Pencil sharpeners are mechanisms for pencil sharpening (and for the production of pencil shavings, and of the distinctive grinding noise that occurs during pencil sharpening, *&c.*). Automobiles contain mech-

[13]This example, and the following are due to Glennan (personal communication).

anisms that explain acceleration and so forth. Homeostatic mechanisms, insofar as they are explanatory, explain a lack of change, *stasis*.

That the room is a particular temperature may not require explanation; if the temperature outside the room is 20°C, then the room would have been 20°C anyway, and the thermostat is playing absolutely no role in maintaining this comfortable state and therefore is not part of the explanation for the room's temperature. That the room is a comfortable 20°C, given that it is very cold outside, does require explanation, because then the thermostat is engaged in a kind of active process of staving off the cold, even as the temperature of the room itself remains static. But notice, in this case, there is a cause and effect relationship that is being explained by appeal to the thermostat: How is it that the thermostat prevents deviations in temperature outside the room from affecting the temperature inside? The cause is the cold temperature outside; the effect is warm temperature inside. And the thermostat is the mechanism linking them: It senses slight drops in temperature due to the outside cold, and brings the heater online to warm the air back up.

Here, then, is a clear sense in which homeostatic machines can be described as etiological mechanisms.

### Correlative Mechanisms

Although neither is a cause of the other, we might ask how it is that the formation of storms and the falling of barometers are linked (because they are). One can reasonably talk of a mechanism responsible for the correlation: Changing air pressure is a mechanism with two distinct (but merely correlated) effects. In this case, changing air pressure is a mechanism that links two effects as correlates, rather than linking a cause to an effect.

While I grant that this is a perfectly fine use of 'mechanism', as with the previous

examples, it can be seen as an etiological mechanism. Specifically, it can be seen as a pair of overlapping etiological mechanisms, with a change in air pressure as the explanandum cause, and the formation of storms as one explanandum effect, and the falling barometer as the second explanandum effect. The falling pressure produces both through different, if overlapping mechanisms. In the case of the storm, decreases in air pressure cause atmosphere water to rapidly condense, forming clouds and eventually rain. In the case of the barometer, decreases in air pressure cause a column of mercury to fall under its own weight, changing the indicator.

So although in one sense it is true that falling air pressure is the mechanism for the correlation, we can re-describe this system as a pair of etiological mechanisms, one for each of the joint effects.

I do not take this to be an exhaustive list of the kinds of non-etiological mechanisms, but rather a demonstration of how such mechanisms can be re-described as etiological mechanisms.

### Etiological Mechanisms and Relevance

Mechanisms link an explanandum cause to an explanandum effect as cause to effect. If this etiological view is correct, how then can we mine it for a principle of mechanistic relevance?

The etiological view offers a ready way to identify an active component as mechanistically relevant: The mechanistically relevant active components are all and only those active components that cause (perhaps indirectly) the explanandum effect, and that are caused by (perhaps indirectly) the explanandum cause. Let me defend this claim.

Figure 4.6 illustrates a causal system with many active components. Craver (2007) calls variables such as $S_1$ and $S_2$ in the figure that so not cause the explanandum

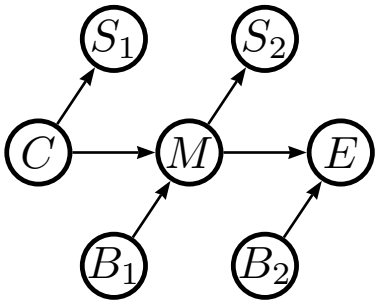


Figure 4.6: $S_1$ and $S_2$ are a sterile effects; $B_1$ and $B_2$ are background conditions; none but $M$ properly belong to the mechanism linking $C$ to $E$.

effect *sterile effects*—and notes that while they may be very useful in identifying the components of a mechanism, they are not themselves explanatory and hence cannot be mechanistically relevant, precisely because they do not in any way connect up with the explanandum effect, and hence cannot *link* the explanandum cause to the explanandum effect. Depending on the context, components such as $B_1$ and $B_2$ in the figure that are not caused by the explanandum cause might be called background conditions, or conflaters, or simply irrelevant. None of these categories are relevant to a mechanistic explanation either, for the same reasons: They are independent of the explanandum cause, and cannot, therefore, illuminate the causal link in question. Only $M$ in the figure properly belongs to the mechanism linking $C$ to $E$, because only $M$ is both caused by the explanandum cause $C$, *and* causes the explanandum effect $E$, therefore linking $C$ and $E$ as cause to effect.

Let us formalize this principle, and see what it gets us.

## M-separation

I have argued now that a mechanism links an explanandum cause to an explanandum effect. This etiological conception immediately suggests a flat (non-hierarchical) representation of a mechanism using graphical models: If $C$ represents the explanandum

cause, $E$ the explanandum effect, and $M$ the mechanism that links them, then an explanation of the causal relation $C \to E$ should have the general form $C \to M \to E$. But, there is much work to be done: Mechanisms contain, in general, many parts, and many causal relations. Thus, in all but the simplest mechanistic explanations, the actual mechanism is something more complex, and we need a clear method for articulating when a component is and is not a part of a mechanism.

Suppose that we have a set of candidate mechanism components, and we wish to know: Which are part of the mechanism linking $C$ to $E$? On an etiological view, since the mechanism links the explanandum cause to the explanandum effect *as* cause to effect, then we should be able to trace a directed path from $C$, through $M$, to $E$.[14] We can trace such a path because mechanisms are an effect of the explanandum cause, and a cause of the explanandum effect.[15]

But it is not enough that there be just some path from $C$ through $M$ to $E$. If $C$, $M$, and $E$ are sets of variables, we want each explanandum cause and each mechanism component to be doing some kind of causal work to bring each of the explanandum effects about. Each explanandum cause should be playing some role in setting the mechanism up, and each component of the mechanism should be contributing to the production of one or more of the explanandum effect.

Given this consideration, here is a graphical principle that I call *m-separation*, meant to capture the above intuitions: We will assign a semantics below!

**Graphical M-Separation (GMR)** Given three disjoint sets of variables, $A$, $B$, and

[14]Recall that a directed path from $A$ to $B$ is a path that begins at $A$, whose first link is an arrow that points away from $A$, in which each link points in the same direction, and in which the final link is an arrow that points to $B$. See Chapter 1.

[15]Mechanistic explanation makes a transitivity of causation a fundamental assumption. If it is true that $C$ causes $E$, that there is a mechanism $M$ that explains this causal link, and we accept Salmon's asymmetry thesis that effects do not explain causes, nor do correlates explain each other, then we are committed to the notion that $C \to M \to E$. Thus, we mechanists are committed to the transitivity of causes—if only within mechanisms.

$C$, $B$ m-separates $A$ from $C$ if and only if

1. all directed paths from any member of $A$ to any member of $C$ pass through and only through one or more members of $B$, and
2. there exists at least one such path.

The first condition requires that any causal relationship between a member of $A$ and a member of $C$ is mediated by one or more members of $B$. It also requires that there be no intermediaries between $A$ and $B$, or between $B$ and $C$. That is, if $A \rightarrow x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow C$, then the set comprising $x_1$,$x_2$,$x_3$ m-separates $A$ from $C$, but no proper subset of those three will so m-separate. The second condition simply gives the universal quantifier in the first condition existential import. Otherwise, the first requirement could be trivially satisfied by two variables $A$ and $C$ with no path connecting them.

Refer back to Figure 4.6. In that figure, there is a directed path from $C$ to $E$, $CME$. $M$ and only $M$ lies on this path; None of $S_1$, $S_2$, $B_1$ or $B_2$ lie on this path. Therefore, in this graph, $M$ is said to m-separate $E$ from $C$.


Figure 4.7: Only in structures as this does $B$ m-separate $A$ from $C$ *and* $C$ from $A$.

Notice that m-separation is asymmetric: If $B$ m-separates $A$ from $C$, it need not necessarily m-separate $C$ from $A$. For it to do so, there must be a return path from $C$ to $A$ that passes through $B$, as in Figure 4.7.

We can now easily apply m-separation as a mechanism bounding principle, by applying the semantics for mechanism models developed in the first half of this chapter.

When we interpret variables as active components, and arrows as interactions, we can apply m-separation to yield the following principle:

**Graphical Mechanistic Relevance (GMR)** Given a set of explanandum causes $C$, a set of explanandum effects $E$, and the smallest set of active components $M$ that m-separates $C$ from $E$, then $X$ is mechanistically relevant in the explanation for $E$ if and only if $X \in M$.

Reading arrows as causal relations, **(GMR)** says that all and only the components in a graph which are caused by one or more of the explanandum causes (directly or indirectly) *and* cause one or more of the explanandum effects are the parts of the mechanism linking the two explananda sets. This is precisely what we wanted.

Notice that **(GMR)**, by explicitly harnessing the interactivity view of mechanism model semantics, provides a bridge between graphical models and mechanisms. **(GMR)** uses m-separation to provide an account of mechanistic relevance, in the sense of being a principle for identifying which active components are relevant to a mechanistic explanation.

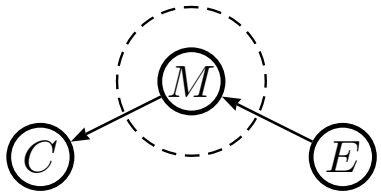


Figure 4.8: Mechanistic relevance is an asymmetric relation. Here, $M$ is not a mechanism linking $C$ to $E$ as cause to effect.

A crucial point to note is that **(GMR)** demonstrates not all causal structures count as mechanisms. Consider the causal structures in Figures 4.8 and 4.9. In neither case does $M$ count as a mechanism linking $C$ to $E$, because in neither case does $M$

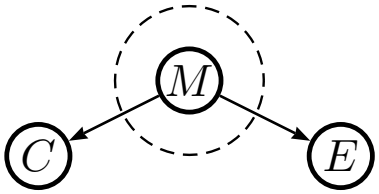


Figure 4.9: Common causes are not mechanisms.

m-separate $C$ from $E$. In Figure 4.8, $M$ m-separates $E$ from $C$, but not $C$ from $E$, and hence cannot be the mechanism linking $C$ to $E$ as cause to effect—because $C$ is not related to $E$ as cause to effect at all.

From the structural constraints imposed on mechanism models by the descriptive accounts of mechanism, I've developed an etiological bounding principle, **(GMR)**, that can tell us whether a given component is a part of a mechanism by examining the arrows in the graph, and hence constrains the range of possible causal structures a mechanistic explanation can have. What **(GMR)** doesn't give us, unlike Craver's **(MM)**, is a testing procedure. For that, we can turn to the Interactivity View's semantics for arrows, that they represent not just unanalyzed causal relations, but relations of manipulability.

The manipulationist semantics gives us a ready testing procedure for **(GMR)**. If an arrow from $A$ to $B$ means that we can manipulate $B$ by intervening into $A$, then the Interactivity View yields the following:

**Etiological Mutual Manipulability (EMM)** An activity-component pair $X$ is part of a mechanism linking a set of explanandum causes $C$ to a set of explanandum effects $E$ if an only if

(i) we can manipulate $X$ by intervening into one or more of $C$, and

(ii) we can manipulate one or more of $E$ by intervening into $X$.

On the Interactivity View semantics, Etiological Mutual Manipulability[16] will reveal which variables will satisfy **(GMR)**, in that observed relations of manipulability must be represented by unidirectional directed paths (perhaps of length one) in a model of the tested mechanism. Because **(EMM)** yields components caused by the explanandum cause, and that cause the explanandum effect, components that satisfy **(EMM)** will be in any set that m-separates the explanandum cause from the explanandum effect.

Of course, not every causal relation will be revealed by ideal interventions, even if they do support manipulationist counter-factuals. For example, if $X$ causes $Y$ and $Z$, and $Y$ also causes $Z$ so as to exactly counter-act $X$'s causal influence on $Z$, then even ideal interventions on $X$ will not reveals the $X \rightarrow Z$ link. This means that **(GMR)** does not entail **(EMM)**, at least not without the additional (and obviously contentious) assumption that there are no such counteracting relationships. However, this is not a problem for my view specifically, but a general shortcoming of any manipulationist view. The upshot is that we can still judge as might be right that $X$ is mechanistically relevant, yet unable to test the claim empirically.

The Interactivity View and Etiological Mutual Manipulability explicitly link the descriptive components of mechanistic explanation—activities and entities—with the formal apparatus of manipulationism and graphical models. Given a particular mechanism, or a description of a mechanism, these principles constrain the range of possible causal models that can be used to represent that mechanism, and by connecting these models to manipulationism, yield a method for drawing inferences about the mechanisms or the descriptions.

I turn now to put the my account of mechanistic relevance, and of the semantics

[16]Of course, there is no mutual manipulability on this view, but it does share what I think is an interesting affinity to Craver's (2007) account of mechanistic relevance. The name is meant to express this affinity.

for mechanism models, to test by applying them to real mechanisms, and showing that the resulting rapprochement is descriptively adequate.

## 4.3 The Manipulated Mechanism

The Interactivity View and Mechanistic Relevance are two more steps forward for the rapprochement begun by Glennan (2002); Woodward (2002) and Craver (2007). These two principles specify how to bridge the qualitative, descriptive elements of mechanistic explanation with the formal elements of manipulationism and causal modeling. I call the resulting rapprochement *the Manipulated Mechanism*.

Of particular interest, the Manipulated Mechanism has descriptive adequacy missing from earlier versions of rapprochement. To demonstrate, I apply the Manipulated Mechanism to two examples: the planetary pencil sharpener, and the production of the action potential in the squid giant axon. In each case I will work from a description of the mechanism to show how my account can be used to build up an interpreted causal model that captures the central qualitative aspects of each mechanism.

### The Pencil Sharpener

The planetary pencil sharpener is a staple of academic life, and an archetypal machine (in Glennan's sense of the term). To use it, one inserts a pencil into the hole while turning a crank; in due course the pencil is neatly sharpened. We may request a mechanistic explanation of it: How is it that turning the crank yields a sharp pencil?

Figure 4.10 reveals the internal workings of the planetary pencil sharpener. The key to the pencil sharpener is a system of planetary gears (from which the style of sharpener gets its name). The crank is bolted to a gear carrier, and turning the crank causes this carrier to rotate about the pencil axis. Contained within the gear

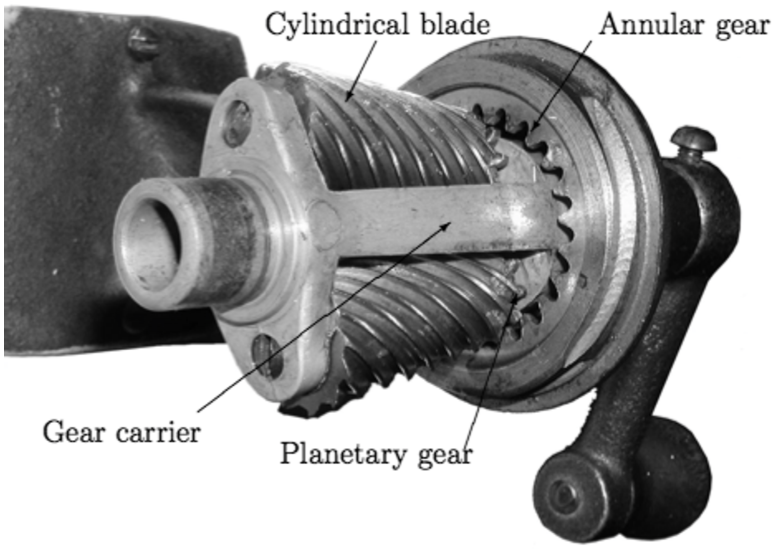


Figure 4.10: Planetary pencil sharpener internals.[17]

carrier are two planetary gears that revolve around the pencil. The planetary gears engage a fixed annular gear which causes the planetary gears to counter-rotate, that is, to rotate against the direction of their revolution. Finally, attached directly to the planetary gears are two cylindrical blades held at an angle within the gear carrier. These blades are made to revolve around the pencil while counter-rotating against the pencil as they revolve. These two motions cause the blades to shave off the pencil's material in a uniformly circular manner.

The first step in constructing a model of this mechanism is to identify the explanandum cause and explanandum effect. Given the rather explicit request just above, the choice is clear: The explanandum cause is the crank's turning ($C$), and the explanandum effect is the pencil's being sharpened ($S$).

The second step is to identify the remaining activity-component pairs. The gear carrier rotates ($G$). The planetary gears revolve ($P_1^r, P_2^r$) and counter-rotate ($P_1^c, P_2^c$).

[17]Photo credit: Toytoy.
`http://commons.wikimedia.org/wiki/File:Pencil_sharpener_mechanism.jpg`.

The annular gear, being fixed, opposes the revolution of the planetary gears which sets them counter-rotating ($A$). Finally, the cylindrical blades revolve around the pencil ($B_1^r, B_2^r$), and counter-rotate against it ($B_1^c, B_2^c$).

The third step is to identify the interactions among the activity-component pairs. The crank rotates the gear carrier ($C \rightarrow G$). The gear carrier revolves the two planetary gears around the pencil ($G \rightarrow P_1^r$ and $G \rightarrow P_2^r$). The revolution of each planetary gear is opposed by the annular gear, which forces each planetary gear into counter-rotation ($P_1^r \rightarrow P_1^c$ and $A \rightarrow P_1^c$; ditto for the second planetary gear). Since both planetary gears are fixed to the cylindrical blades, the rotation and revolution of each gear forces the cylindrical blades to revolve and counter-rotate as well ($P_1^r \rightarrow B_1^r$ and $P_1^c \rightarrow B_1^c$; ditto for the second cylindrical blade). Finally, the revolving and counter-rotating of the cylindrical blades shaves material off of the pencil ($B_1^r \rightarrow S$ and $B_0^r \rightarrow S$; ditto for the second cylindrical blade).

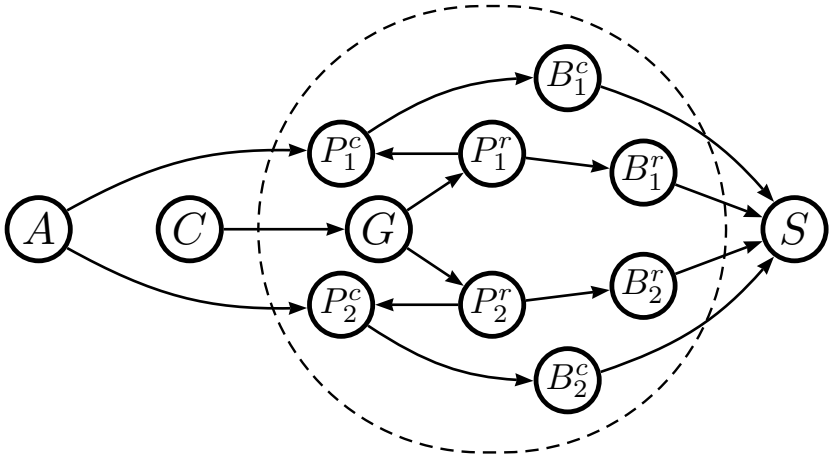


Figure 4.11: A causal model of the mechanism connecting the crank to pencil sharpening.

The final step is to put all of these parts together into a graph. Figure 4.11

shows the results. Applying **(GMR)** to the resulting graph reveals the mechanistically relevant components (indicated by the dashed outline in the figure). All and only $\{G, P_1^r, P_2^r, P_1^c, P_2^c, B_1^r, B_2^r, B_1^c, B_2^c\}$ are on a directed path from $C$ to $S$, and thus count as mechanistically relevant components.

Interestingly, $A$, the annular gear's opposition to the planetary gears' movement, is not mechanistically relevant according to **(GMR)**, and thus is not a proper part of the mechanism. But this result is at odds with our intuitions that the annular gear is very much a component of the pencil sharpener. This result brings into sharp relief the distinction between our colloquial notions of mechanism or machine, and the philosophical concept of mechanistic explanation. In an everyday sense of the word 'mechanism', of course the annular gear is a part. But so too, on this view, are force by which the crank is turned, the casing, the wall-mounting, and the art-deco trim. Such use of the term 'mechanism' is meant to pick out an entire, discrete *object*.

But a mechanistic explanation, in contrast, seeks to understand the causal links between a stipulated cause and effect, and so some parts of the pencil sharpener will naturally not be parts of the mechanistic explanation (and perhaps some components *not* part of the pencil sharpener might turn out to be mechanistically relevant). The art-deco trim, for instance, clearly plays no causal role in bringing about a sharp pencil, and so it does not belong to the mechanism in this sense of the word. But the annular gear *is*, one might object, causally relevant to pencil sharpening. And indeed it is—but this is where mechanistic explanation parts ways with more generally causal explanations. Crank-turning is not causally relevant to the working of the annular gear; the annular gear is not in any sense causally between the explanandum cause and explanandum effect, and thus is properly left out of a mechanistic explanation.

The point can be considered in a different way. Suppose that we held the crank fixed, and rotated the annular gear only. If we did, we would find that we could

sharpen a pencil in this way, and we might well ask: What is the mechanism linking the rotation of the annular gear to pencil sharpening? And the result would be much the same as in Figure 4.11, save that $G$ would be excluded as mechanistically irrelevant. Thus, there are two distinct ways to use a pencil-sharpener to sharpen a pencil, and two different (but overlapping) mechanisms that describe these two ways.

Here, the distinction between colloquial notions and philosophical notions of mechanism becomes sharpest. Where the colloquial sense recognizes but one thing, the pencil sharpener, the philosophical sense recognizes several partially overlapping things, including the mechanism linking the crank to sharp pencils, and the mechanism linking the annular gear to sharp pencils. There is also the mechanism linking the crank and annular gear taken together to sharp pencils. Too, there is the mechanism linking the crank to the production of pencil shavings, the internal clockwork's inefficiency to our hearing of pencil-sharpening-noises, the art-deco trim's reflectivity to our aesthetic judgments of the pencil sharpener, and so *ad infinitum*. So, within the single object of the pencil sharpener reside many, perhaps infinitely many, mechanisms, depending on how one picks out the explanandum cause and the explanandum effect. But this is just a consequence of the idea that, as Craver puts it, 'there are no mechanisms simpliciter; all mechanisms are mechanisms of something, and it is by reference to that something that the relevance of components is established.' (Craver, 2004, p. 969)[18]

So much for my toy example. Let us turn to some biology and see how the manipulated mechanism fares there.

[18]The annular gear in this example is meant to be analogous to a classic example of a background condition, oxygen. We might ask: What caused the fire at the factory? and accept: The arsonist's torch. as a suitable response, and yet reject: The influx of oxygen. as inappropriate. The reason is that, in this context, oxygen is a background condition. Although it is true that the fire would have been short-lived without oxygen, it is irrelevant to the desired explanation. In the same way, the annular gear plays a critical role in the sharpening of pencils, but in the context of the explanation requested, it is a background condition. I take this affinity to be a strength of my view, as counter-intuitive as the result might otherwise be.

## Neuron Depolarization

The action potential is an electrical pulse that is transmitted down the axon of neurons. It has the interesting feature that, once started, it cannot be stopped or attenuated (without external disturbance, anyway). Machamer, Darden, & Craver (2000) use the mechanism for the action potential as a test case for their account; I shall follow suit.

The action potential has three stages: As the action potential reaches a region of the axon, it triggers a *depolarization*, a rapid upward change in the transmembrane voltage from a roughly -70mV resting state. Then, as the action potential passes, it triggers a *hyperpolarization* wherein the transmembrane voltage is driven back down well below the resting state. A *refractory period* follows during which resting state equilibrium is restored. I discuss here the mechanism for depolarization.

Embedded within the cell membrane a a number of voltage-gated sodium ($Na^+$) channels. These channels are composed on one large protein with four identical parts. These parts are elongated, and span the membrane; they are grouped close together so as to form a pore. Each of the four parts consists in two domains: A voltage sensor and a gate.

The voltage sensor is constructed of four $\alpha$ helices[19], of which one (S4) has a regular sequence of positive charges along its length. The remaining three have negative charges along their length. At rest, the negative charges and the positive charges align, and the negative charges hold the S4 helix in equilibrium. An approaching action potential, however, disrupts this balance: The changing voltage gradient across the membrane disrupts the equilibrium, and S4 rotates such that the positive charges come out of alignment with the negative charges, and so is pushed out of the cell like

[19] An $\alpha$ helix is a kind of protein structure in which a length of protein is twisted into a helix at the rate of 3.4 acid bases per turn.

a corkscrew (Hall, 1992).

The relationship between the voltage sensor and the gate is still something of a mystery. The gate domain contains two helices, S5 and S6; normally, S6 blocks the channel. But the movement of the S4 helix induces the S6 helix in the gate to bend in the middle, opening the channel to the flow of sodium (Sands, Grottesi, & Sansom, 2005).

As the sodium channels open, sodium is pushed out of the cell by the electromotive force of the approaching action potential. The resulting change in relative sodium density across the cell membrane results in a local depolarization, which in turn causes more sodium channels to open, forming a positive feedback loop (Hall, 1992). In this way the leading edge of the action potential is propagated: As depolarization increases at one locale, it spreads as a weaker depolarization further down, triggering sodium channels there to open, and so forth all the way down the axon.

First, we identify the explanandum. The explanandum causes are the approach of the leading edge of the action potential and the initial resting conditions (the equilibrium sodium concentration; the location and orientation of the ion channels, etc.). The explanandum effects are the radical local depolarization (the peak of the action potential, and its movement) and the increase in the intracellular concentrations of sodium.

Second, the entities and activities. The first is the relative charge across the cell membrane, the membrane potential ($V$). The voltage sensor S4 detects changes in the trans-membrane voltage by providing a charge to be impelled; hence the S4 can move outwards from the cell membrane ($m$). The gate S6 can undergo a conformance change, bending like a straw, opening and closing the channel and making it more or less conductive to sodium ($g$). Finally, sodium ions can flow across the membrane in

greater or lesser quantity ($I$).[20]

Third, the interactions. The membrane potential impels the voltage sensor slightly out of the cell ($V \to m$). The movement of S4 triggers a conformance change in S6 ($m \to g$). The conformance change opens the channel to the movement of sodium ions ($g \to I$). The membrane potential pushes the sodium ions through the open channels ($V \to I$). Finally, the movement of sodium ions directly alters the membrane potential ($I \to V$).

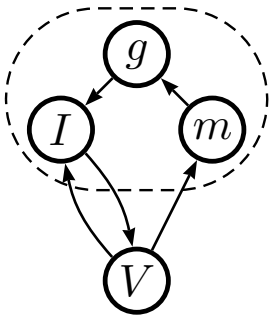


Figure 4.12: A model of the mechanism of axon depolarization during an action potential.

These interactions are combined in the model of Figure 4.12. Notice that we identified particular states of the membrane potential as both explanandum cause and explanandum effect: We asked, how is it that a slight upward disturbance in the membrane potential can rapidly and radically amplify itself? Thus, we are asking for the mechanism for a self-triggering positive feedback mechanism—we should not surprised to see that $A$ is both explanandum cause and explanandum effect.

Not in the model, however, are the initial conditions nor the final conditions. The location and orientation of the ion channels is implied by the causal dependencies in the graph: If the orientation or location were otherwise, then the the causal de-

[20]The variables were chosen to align with the choices of Hodgkin & Huxley (1952d).

pendencies exhibited by the axon would be quite different. But Machamer, Darden, & Craver also specify the initial states of the activity-component pairs as part of the explanandum cause. Representing these initial and final states requires a functional model. Graphical models cannot themselves represent the functional relations among activity-component pairs—they can only demonstrate that such relations exist. To complete the model, we require a functional model that spells out the observed mathematical relationship among the active entities during experimental interventions. Once we have done so, we will have accounted for the initial and final states of the activity-component pairs. Indeed, we will find that Hodgkin & Huxley (1952d, pp. 504–507,512) have already done this work for us with their mathematical model relating the membrane potential to ion current:[21]

$$\begin{aligned} I &= gV \\ g &= m^3 h\bar{g} \\ \frac{dm}{dt} &\propto V \end{aligned}$$

The first equation claims that the sodium current is proportional to the sodium channel conductance and the deviation of the membrane potential from rest; the second that the channel conductance is proportional to the third power of $m$, which Hodgkin and Huxley interpreted as representing what we now call the gating charge—effectively a representation of the outward movement of S4; and $h$ which represents gate *inactivation* (which is not part of the mechanism for depolarization, but for the later hyperpolarization), and a constant. The third entry above summarizes Hodgkin and Huxley's demonstration in a sequence of equations that $m$ is a function of and proportional to the membrane potential.[22]

[21]The Hodgkin and Huxley model is of course comprehensive over sodium *and* potassium currents, as well as leakage, capacitance, and other electrical currents. I focus here on only the part of the model dealing with sodium, and so leave off the relevant subscripts.

[22]The change in $m$ is a complex function of $V$. I have summarized the function by simply noting

To this, we need only add Ohm's law:

$$V = I\frac{1}{g}$$

which claims that voltage is proportional to current and inversely proportional to conductance. Given causal interpretations, in which the variable on the left-hand side is read to be an effect of the variables on the right-hand side, these equations serve as a functional adjunct to the graphical model in Figure 4.12. Notice, in particular, that although I have derived the graphical model from a narrative description of the mechanism, we could have derived just the same model directly from this functional model, once given a causal interpretation. This kind of convergence speaks to the descriptive *and* quantitative adequacy of my account of the manipulated mechanism

## 4.4 Conclusion

Formal systems are, by their very nature, uninterpreted: Rich in syntax but devoid of semantics. But accounts of mechanism are not, nor can they be, purely formal enterprises. A semantics is thus required to apply causal models to the job of mechanistic explanation. A complete quantitative account of mechanism should bring together the syntax of a formal framework for causal reasoning, and the semantics of qualitative accounts of mechanisms. Previous work on the rapprochement have faltered either because they have unfettered themselves from the qualitative constraints of mechanistic explanation, or the formal elements of manipulationism. In this chapter, I have considered a two new elements for the rapprochement that neglect neither the syntax nor the semantics, the Interactivity View, and Graphical M-Separation (with its testing procedure Etiological Mutual Manipulability). The Manipulated Mechanism, as I call my account, retains the descriptive strength of Machamer, Darden, &

that the change in $m$ is proportional to $V$ to avoid bogging the discussion down in arcane details.

Craver (2000) and Glennan (1996) and the prescriptive strength of Woodward (2003). In this way, manipulationism and causal modeling open the door to a wide range of prescriptive and normative enterprises: evaluating mechanistic explanations, judging when a component is or is not part of a mechanism and hence its relevance to an explanation, and using experimentation to discover mechanisms.

I turn now to defend the rapprochement developed here against a recent and compelling objection raised against Woodward's manipulationism: That the rapprochement require mechanisms to be *modular*, where in fact very few mechanisms actually are.

# Chapter 5

# The Principle of Modularity

My account of the manipulated mechanism is founded on Woodward's manipulationist account of causal relevance. My account, therefore, inherits the assumptions of manipulationism. Central among the assumptions of manipulationism is the principle of *modularity*, which is (roughly) the assumption that we can intervene into the various parts of a mechanism independently of each other. Modularity faces a number of challenges, but in my opinion they each rest on a failure to satisfactorily pin down the idea motivating modularity. Some turn on intuitive notions orthogonal to the intuitions behind modularity. Some include in the concept of modularity additional very strong assumptions. In this chapter, I set up a defense of modularity by trying to be explicit about this central bridge principle, to understand the causal intuitions driving it, to lay out its various formulations, and to lay out its working parts so that we can know which objections strike at its heart, and which are irrelevant to our ability to make inferences on the basis of causal interventions. By the end of the chapter, I will defend Woodward's Probabilistic Modularity as the right articulation of modularity to found a defense of modularity.[1]

[1]With thanks to Craver (personal communication) for helping articulate the goal of this chapter.

I perform an experiment, and conclude from the resulting data that the independent variable must be a cause of the dependent variable.[2] But to draw this conclusion I need an assumption or assumptions that permit me to move from a set of observed correlations to a causal claim—after all, correlation does not (alone) imply causation.In this chapter, I explore the assumptions necessary for this kind of inference, and focus my attention on one controversial assumption known as modularity.

In §5.1, I examine why we might think experimental manipulation is a good form of causal inference. I begin by articulating what I take to be a basic intuition about how causal relations behave during an experimental manipulation, namely that the correlations generated by causal connections are *robust* in a way that merely accidental correlations are not, and that therefore causal correlations will survive at least some experimental manipulation where accidental correlations will not. This intuition gives rise to two assumptions. The first assumption, which I call the *asymmetry of causal correlation* **(ACC)** is that causal correlations will survive experimental manipulation of the cause only; the second assumption, which I call *the symmetry of accidental correlation* **(SAC)** states that accidental correlations will not survive experimental manipulation of either correlate. I argue that **(SAC)** is the most basic articulation of modularity: That interventions should be able to break the correlations linking the intervened-upon entity from its non-effects.

But, **(ACC)** and **(SAC)** are not themselves useful for causal inference in the real world. Instead, we need a modularity principle that builds upon **(SAC)**, that can connect our intuitions about causal correlations to formal systems of inference. Many

[2]Of course, our conclusions will be more carefully couched than this simple gloss. But this is nevertheless the kind of positive conclusion I would *like* to be able to make, and so I run with it. The couching and qualifying of biologists' *actual* conclusions largely stem from difficulties in the analysis of the experimental data (and a general conservative hesitance). Statistical methods require assumptions that can't always be guaranteed, and the results of statistical analysis are always probabilistic in nature. Assuming that statistical methods can return reliable, incontrovertible results is obviously a gross idealization on my part, but it gets the ball rolling as a limiting case.

authors (Fodor, 1983; Cooley & Leroy, 1985; Hausman & Woodward, 1999) have offered a broad range of more nuanced modularity principles; a somewhat confusing accident of the literature is that each has called their own version "modularity". But each author augments **(SAC)** in a different way, and to a different degree. In §5.2, I lay out some of the most well-known versions of modularity, show how each is in some sense analogous to **(SAC)**, and how each varies from either the intuitions behind or the inferential purpose of **(SAC)**. Most of these principles have too-specific subjects to be of general interest, or do not in themselves license causal inference from experimental manipulation. As such, these principles are not workable for purposes of manipulationist causal inference, and are not appropriate targets for the critics of modularity.

On the other hand, Woodward's (2003) *Probabilistic Modularity* **(PM)** was formulated specifically for experimental inference and is the centerpiece of his manipulationist framework—and by extension of the manipulated mechanism as well. In §5.3, I argue that it offers a bridge principle that links experimental observation with causal inference, unlike the principles considered in the previous section, and is therefore the right articulation to defend. I argue that **(PM)** is a workable modularity principle that describes our modular intuitions, and licenses causal inferences from experimental manipulation in the same way as **(SAC)**. My argument proceeds by analyzing Woodward's probabilistic modularity into two distinct principles, which I helpfully label **(PMa)** and **(PMb)**. **(PMa)** bears a strong resemblance to **(SAC)**, but **(PMb)** is not itself a modularity principle.

In the following chapter, I will show that some objections to modularity are really objections to the non-modularity principle **(PMb)** contained within **(PM)**, and that we can avoid these objections by jettisoning **(PMb)**. Other objections work as objections to **(PMa)**, however. In the final two chapters, I will show that these objections

rest on a misconstrual of the metaphysical demands of modularity. At any rate, my goal in this chapter is to get maximally clear as to what modularity is, to show how various articulations of it relate to each other, and to set up the defense of modularity executed in following three chapters.

## 5.1 Modular Intuitions

Before we can select a formulation of modularity to defend, let us step back to consider what a modularity principle is, and why we need one for inference from experimental manipulation. I begin by examining two basic intuitions about causal relations, and derive from them two basic principles **(ACC)** and **(SAC)**. **(SAC)** is of particular interest, as it is the most basic formulation of a modularity principle; getting clear on the role that **(SAC)** plays in causal inference in toy systems will serve as a useful foundation for evaluating the more nuanced modularity principles I examine in the next section.

We often think that when one thing causes another, that the two will exhibit some correlation. Indeed, this is likely the only direct evidence we could possibly have for a causal relation. But correlation alone is not enough to infer causation, because there are many non-causal reasons why two things might be correlated (see Haig, 2003, for an overview of different kinds of non-causal correlation). Accidental correlations are profligate: As Sober (2001) famously observed, the price of bread is closely correlated with the sea level in Venice. Accidental correlations can also arise from a common cause. Thus, we need a procedure for filtering causal correlation—the kind that arises when one correlate is the cause of the other—from these other kinds of correlation (which I will lump together under the heading of 'accidental correlations').

Biologists have long considered experimental manipulation a useful method for

filtering accidental correlations from causal correlations. In general, experimental inference works something like this: When I manipulate $A$, and I observe that $B$ is correlated with my intervention into $A$, then I conclude that my manipulation of $A$ *brought about* the changes in $B$—that $A$ causes $B$. The presumption is that there is something about the act of intervention that permits us to conclude that the correlation between $A$ and $B$ is not accidental. This assumption begins to find modern form in Mill's (1843) Methods, and lies at the heart of accounts by von Wright (1971), Menzies & Price (1993), Pearl (2000), and Woodward (2003).[3]

The experimental mode of inference rests on two assumptions that revolve around intuitions that *experimental interventions are somehow special*, in that whatever correlates with an intervention must have been caused by that intervention. Put slightly differently, experimental inference relies on the idea that accidental correlations are fragile and do not survive experimental manipulation, where causal correlations are robust and do survive experimental manipulation.

The first intuition is that causes are difference makers. Introducing a change into a cause will introduce changes into its effects[4], and these changes will be correlated (when they occur). Building on this intuition, the first assumption is that when one thing[5] is a cause of a second and the two are correlated, then the correlation between the two will survive interventions into the first, the cause. Call this assumption *the asymmetry of causal correlation* or **(ACC)**.

[3]Of course, the causal Markov condition (described in Section 1.1) is another method for sorting correlations that arise from common causes from directly causal correlations (although it requires an assumption that other kinds of accidental correlation do not arise). I consider the relationship between the causal Markov condition, and the principles underlying experimental manipulation in Chapter 6.

[4]Roughly. I am explicating an intuition, not offering a rigorous account of causation. I, for now, ignore issues about cancelation of effects, transitivity, and other ways that wiggling a cause might not lead to a wiggle in the effect.

[5]I use the deliberately ambiguous term 'thing' here as a placeholder for whatever we decide causal relata to be; here we are after a sign of the causal relation, and so the nature of the relata doesn't yet concern us.

The second intuition is that accidental correlations are not a channel through which we can exert control; this is a hallmark of causal relations only. Introducing a change into an accidental correlate will *not* introduce changes in the other correlate. Building on this intuition, the second assumption is that when one thing is *not* a cause of a second and the two are correlated, then the correlation between the two will disappear under interventions into the first. Call this assumption *the symmetry of accidental correlation* or **(SAC)**.

**(SAC)** is a *modularity principle*. It captures the important if perhaps less common intuition that causes are the kind of thing that we can get ahold of independently of each other—that causes are, in some vague sense, modules. Suppose we have two (causally unrelated) causes before us: Our intuition is that the way we manipulate one of the causes places no constraints on how we can manipulate the second cause. If these two are not causally related, **(SAC)** tells us that any correlation between them will disappear once we intervene into either one of them. But, if the manipulation of one created constraints on how we could manipulate the second, these constraints would appear as a correlation, an outcome that **(SAC)** specifically rules out. To offer a simplistic illustration, suppose I have before me two knobs. There's no reason why I couldn't turn both of them this way and that in precisely the same way. But I could also rotate each of the two knob any way I choose, without one knob being constrained by the movement of the other. I can manipulate them independently of each other.

**(ACC)** and **(SAC)** together form the foundation of a (rather primitive) system of causal inference from experimental manipulation. This foundation, as we shall see, is very weak: To accomplish any kind of inferential heavy-lifting, **(ACC)** and **(SAC)** will have to be strengthened considerably, and they must be supplemented with additional assumptions about the nature of interventions. But **(ACC)** and **(SAC)** will

work in simple toy systems; and insofar as they get the very simplest cases right, then **(ACC)** and **(SAC)** are good building blocks to start from.

To provide a simple illustration of how **(ACC)** and **(SAC)** work, suppose again a system comprising two variables $A$ and $B$. In such a system there are four possible causal relations among $A$ and $B$: $A$ can cause $B$, $B$ can cause $A$, they can both cause each other, or neither is a cause of the other. Moreover, assume that $A$ correlates with $B$. The question is, given given the correlational data resulting from a pair of interventions—one on $A$ and one on $B$, how can we derive the cause structure of the system?

Table 5.1: Using **(ACC)** and **(SAC)** to map experimental correlations to causal structure.

| | $A \not\leftrightarrow B$ | $A \rightarrow B$ | $A \leftarrow B$ | $A \leftrightarrow B$ |
|---|---|---|---|---|
| $\text{set}(A) \perp B$ | ✓ | | ✓ | |
| $\text{set}(A) \not\perp B$ | | ✓ | | ✓ |
| $A \perp \text{set}(B)$ | ✓ | ✓ | | |
| $A \not\perp \text{set}(B)$ | | | ✓ | ✓ |

**(ACC)**, recall, says that causal correlations survive interventions into the cause, but not the effect. **(SAC)** says that accidental correlations do not survive interventions into either relata. So, if an intervention into $A$ does not correlate with $B$, then **(ACC)** permits us to conclude that $A$ must not be a cause of $B$; if an intervention into $B$ does not correlate with $A$, then **(ACC)** permits us to conclude that $B$ must not be a cause of $A$. Likewise, of an intervention into $A$ *does* correlate with $B$, then **(SAC)** permits us to conclude that $A$ *is* a cause a $B$; if an intervention into $B$ does correlate with $A$, then **(SAC)** permits us to conclude that $B$ is a cause of $A$. Table 5.1 summarizes these results; along the top of the table are the possible causal con-

nections among $A$ and $B$; along the left are the four possible experimental outcomes (with 'set($A$)' indicating an intervention into $A$, '$\not\perp$' indicating correlation, and '$\perp$' indicating a lack of correlation). The check-marks in the table indicate, according to **(SAC)** and **(ACC)**, which set of experimental correlations match each possible causal structure.

Thus, at least in this toy system, having a modularity principle (namely, **(SAC)**) is necessary for causal inference from experimental manipulation (in conjunction with **(ACC)**, of course). The centrality of **(SAC)** in causal reasoning can be brought out by considering failures of the principle.

## Failures of (SAC) and (ACC)

The conditions under which a principle is violated can be instructive for assessing the principle's usefulness.

Violations of **(ACC)** lead us to falsely classify causal correlations as accidental. **(ACC)** says that if $A$ and $B$ are correlated, and $A$ is a cause of $B$, then the correlation will survive interventions into $A$. Thus, **(ACC)** is violated when $A$ and $B$ are correlated, $A$ is a cause of $B$, and the correlation disappears under interventions into $A$. Under these conditions, we could not conclude (on the basis of **(ACC)** alone) that $A$ was not a cause of $B$. Violations of **(ACC)** deny us causal knowledge, but do not lead us to falsely posit a causal connection where non exists. If we take as a reasonable default view that we should assume two things[6] are not causally related until we have positive evidence otherwise, then although violations of **(ACC)** will lead us astray, they will leave us with overly conservative causal models.

Violations of **(SAC)** lead us to falsely classify accidental correlations as causal. **(SAC)** says that if $A$ and $B$ are correlated, and neither $A$ nor $B$ is a cause of

[6]...events, properties, what have you...

the other, then the correlation will disappear under interventions into either. Thus, **(SAC)** is violated when $A$ and $B$ are correlated, neither is a cause of the other, and the correlation remains during an intervention into one or the other. Under these conditions, we would conclude (on the basis of **(SAC)** alone) that $A$ was a cause of $B$. Violations of **(SAC)** make causal connections appear to be more common than they in fact are. Again, if we take as a default view that we should assume two things are not causally related until we have positive evidence otherwise, then violations of **(SAC)** provide false positive evidence, and leave us with causal models with too many causal connections.

Violations of **(ACC)** leave us lacking in causal knowledge, which is certainly a loss. So long as **(ACC)** isn't violated in every instance, we can still learn something true about the causal structure of the world—we are left with no doubts about the causal links we do include in our models. But violations of **(SAC)** leave us with too many causal connections. Even if **(SAC)** is violated occasionally, we *are* left with doubts about the causal links we include in our models. Thus, if we have independent reason to think that **(SAC)** might be violated, then we also have reason to doubt the causal models that we construct using it, and hence the entire system of inference that relies on it. If we were to discover that **(ACC)** were frequently violated, we might wonder what our causal models of the world were missing, but such violations would not cast doubt on previous inferences. On the other hand, if we were to discover that **(SAC)** were frequently violated, such violations *would* cast doubt on previous inferences, for we would have reason to think at least some of them are unjustified. For this reason, violations of **(SAC)** are far more pernicious than violations of **(ACC)**.

The diagnostic conclusion to draw from considering the conditions under which **(SAC)** fails is this; That **(SAC)** is an inferential safeguard that justifies our inference to causal links using **(ACC)** or similar principles. Without a modularity principle

like **(SAC)**, we could not justifiably include any causal link in our models.

So far, we have only considered toy systems comprising two variables. Causal inference in more realistically complex systems requires taking on additional assumptions. I offer **(ACC)** and **(SAC)** as a starting point for thinking about modularity, what it can do for us, and what it requires. **(SAC)** in particular gets at the central idea behind modularity, and my discussion of it in this section is not meant as a weighty treatise on causal reasoning, but as a way of getting as clear as possible about the role that modularity plays in causal inference from experimental manipulation. Thus, having examined the most basic modularity principle, **(SAC)**, I turn now to examine more nuanced modularity principles.

## 5.2 Disambiguating 'Modularity'

A defense of modularity requires a clear statement of the principle to defend. Having considered the overly simple modularity principle, **(SAC)**, I turn now to consider more sophisticated modularity principle formulations: Fodorian modularity, developmental modularity, and Woodward-Hausman modularity.

### Fodorian Modularity

First allow me to clear away a confusion that will likely occur to readers familiar with Fodor's work. Though Fodorian modularity can support certain kinds of causal inference, his idea of modularity is almost entirely distinct from the kind of modularity I am concerned with. Fodorian modularity is a domain-specific, coarse-grained principle, where modularity in the sense discussed in the previous section is a domain-general, fine-grained principle.

Fodorian modularity is domain-specific. Fodor is interested in issues of cognitive,

rather than causal, structure and organization. "Roughly," Fodor tells us, "modular cognitive systems are domain specific, innately specified, hardwired, autonomous, and not assembled"—but they are especially informationally encapsulated (Fodor, 1983). Fodor has crafted his list of attributes to characterize *cognitive* systems, unlike **(SAC)**, which is crafted to describe causal systems generally.

Fodorian modularity is coarse-grained. Fodorian modules are computational, encompassing mechanisms at many levels (in the mechanistic sense of Craver (2007)). To perform one cognitive function will require a great many causal interactions among neurons, each of which will comprise many causal interactions at the molecular level, and so on. Fodorian modularity does not require anything of these lower-level systems; nor does it require anything of higher-level systems comprising many Fodorian modules. In contrast, **(SAC)** is concerned with causal systems at any level of granularity.

That said, Fodorian modularity and **(SAC)** are related, and this relation is worth noting. A Fodorian module is strongly independent of other parts of the mind (modules or not). The interactions—surely a causal relation—between a Fodorian module and other parts of the brain are narrow and restricted—this is the essence of informational encapsulation. What makes a Fodorian module a module at all is that the inner details are utterly opaque to the rest of the system. In this way, the rest of the system, insofar as it depends on a particular module, would not be the least disrupted if that module were replaced with a second module providing the same interface, but comprising radically different internal workings. Thus, we could intervene into a Fodorian module so as to disrupt any accidental correlations with other modules. But where **(SAC)** makes this kind of claim of any causal system, Fodorian modularity makes this claim only of cognitive systems, and only at the level of cognitive modules.

Thus, despite the similarities to **(SAC)**, Fodorian modularity is too narrow for

the purposes of general causal inference. There is no reason to think that *only* informationally encapsulated entities can satisfy modularity, as I use the term. Unlike Fodorian modules, most causal systems do not involve the transmission or transduction of a signal, and when they do, there is no reason to think that signal must be of narrow bandwidth. Genes, as I will discuss below, are often thought to be modular, that is, independent of one another. But, Griffiths (2001) has argued that genes do not encode information—yet the success of Griffiths's argument surely does not undermine the *causal* independence of genes. So, even if we extend Fodorian modularity to cover non-cognitive systems, its central requirement that modules be processors of information is yet too strong for his principle to apply to the kinds of biological cases we are interested in. Thus, when I talk of modularity, I am not speaking of Fodorian modularity.

## Developmental Modularity

Modularity plays a central role in the emerging field of evolutionary-development biology (evo-devo). Seeking to understand the evolutionary history of a phenotype, evo-devo researchers compare the development of related phenotypes across species. The patterns on butterfly and moth wings have provided a fruitful case study for evo-devo researchers. These patterns appear to be composed from genetic building blocks—the pattern elements (*e.g.* coloration, geometry, and repetition) are, in many cases, encoded in distinct genes such that researchers can manipulate one pattern element, leaving the remainder intact. For example, the size of eye-spots on the forewings is independent of the size (or existence) or eye-spots on the hindwings; the coloration of the eye-spot can vary between the forewings and hindwings as well, but these pattern elements cannot be varied between the left and right forewings, or the left and right hindwings. (Beldade & Brakefield, 2002).

These patterns are encoded separately, develop (more or less) independently, and are thus thought to have evolved independently as well. Thus, researchers as Wagner, Pavlicev, & Cheverud (2007) call them modules. "A network of interactions," Wagner, Pavlicev, & Cheverud tell us, "is called [developmentally] modular if it is subdivided into relatively autonomous, internally highly connected components" (Wagner, Pavlicev, & Cheverud, 2007, p. 921).[7] This definition share much in common with **(SAC)**. The general idea is that if a phenotype is developmentally modular, then we can intervene into that phenotype independently of the organism's other phenotypes.

Unlike Fodorian modularity, developmental modularity is not domain-specific: Wagner's definition makes no reference to genes, phenotypes, or any other biology-specific entity. But like Fodorian modularity, and unlike **(SAC)**, it is restricted to fairly coarse-grained systems. On Wagner's view, modularity is a property of a 'network of interactions'; but **(SAC)** is a claim about single interactions. Developmental modularity is a metric for subdividing a complex into simpler parts based on observed features (in a similar manner to Simon's (1962) notion of near-decomposability), where **(SAC)** is a claim about individual correlations during an intervention.

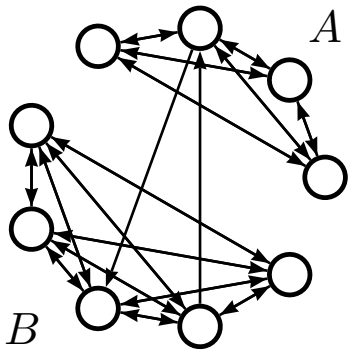


Figure 5.1: A developmental modular system, comprising two modules 'A' and 'B'.

[7] I used the term 'developmental modularity' for purposes of disambiguation only: Nothing hangs on the word choice.

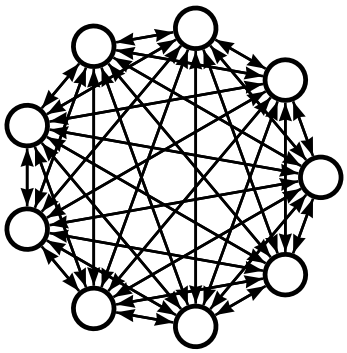

Figure 5.2: A non-developmental modular system.

To illustrate the difference, consider the two causal systems represented in Figures 5.1 and 5.2. The system represented in Figure 5.1 is developmentally modular, because it can be subdivided into two developmental modules, which I have labeled 'A' and 'B', according to the relative density of causal interactions among the nodes in each module: The components are clustered into modules that have more internal connections than external. The system represented in Figure 5.2 is not developmentally modular, because it cannot be subdivided into developmental modules, because it is fully interconnected: There is no subgraph in the system that has more internal connections than external.

Yet **(SAC)** is not a claim about systems, but about individual correlations. An accidental correlation will satisfy **(SAC)** when it disappears under interventions to one or the other correlate. A causal correlation will satisfy **(SAC)** when it disappears under interventions into the effect. Thus, either of the system in Figure 5.1 and 5.2 might contain correlations that satisfy or fail to satisfy **(SAC)**.

Thus, developmental modularity is distinct from the kind of modularity represented by **(SAC)**. Developmental modularity is a metric for clustering systems of interacting entities; the kind of modularity I am after is a principle for discriminating

accidental from causal correlation.

## Woodward-Hausman Modularity

The notion of modularity articulated by Hausman & Woodward (1999, 2004a,b); Woodward (1999, 2002, 2003) is related to, but distinct from the two principles discussed above. Woodward-Hausman modularity[8] is a general principle that places constraints on interventions into a causal system.

> Modularity should be understood as claiming that for each [component] there will be some range of interventions (interventions that are of the "right kind" and not "too big" that disrupt only the relationship between [that component] and its direct causes and no others in the system of interest. We take this to be a substantive claim concerning the nature of causation.... (Hausman & Woodward, 2004a, p. 850)

On their view, modularity is a general, abstract claim that causal inference requires that each component in a causal system be independently manipulable, that is, that we can intervene into each component without disrupting the rest of the system. The worry against which modularity ensures is easy to see: If our interventions *did* disrupt a causal system extensively, then we would not be warranted in drawing any causal conclusions from our intervention. If an intervention into a metabolic pathway utterly destroys it, we would be nevertheless mistaken to then claim that the pathway's precursors cause its destruction.

Notice in particular what Woodward-Hausman modularity *does not* say. It says nothing about what a component is, anything about how richly that component is interconnected with others, how it processes information, or whether it is in any sense

[8] Again, I qualify the label only to ease disambiguation.

opaque or atomic. Woodward-Hausman modularity, unlike Fodorian modularity or developmental modularity, is not a claim about the overall structure of a system, but is a constraint placed on individual causal links within that system. In particular, it requires that if two components are causally correlated, then that correlation will disappear during an intervention into the effect. A system is modular if and only if each component is modular in this way.

In this sense, and again unlike Fodorian or developmental modularity, Hausman-Woodward modularity has a commonality with **(SAC)**, in that both place the same constraints on interventions into the effect of a cause-effect pair. This particular expression of Hausman-Woodward modularity does not, however, make any claims about accidental correlations, and whether they survive or disappear under intervention. As such, it is a principle for determining the direction of causation from interventions into causal correlations, and so can sort accidental correlations arising from common causes from causal correlations, but it does not yet permit us to sort all accidental correlations from causal correlations.

Woodward-Hausman modularity is a good first step towards a general formulation of a modularity principle. Unlike Fodorian or developmental modularity, it captures at least part of the intuition that we can sort different correlations via intervention. I turn now to consider a refinement on Woodward-Hausman modularity, called probabilistic modularity. Probabilistic modularity is of interest because it permits distinguishing purely coincidental correlations from causal relations, and because it provides an explicit bridge between observed correlations and causal models.

## 5.3 Woodward's Probabilistic Modularity

Woodward's (2003) manipulationist view of causal relevance is designed around the idea that experimental intervention is one way to discover causal relations. The concept of modularity, in the sense of a principle for sorting accidental from causal correlations, figures prominently in his account of causal relevance. Woodward provides a refinement on Woodward-Hausman modularity, that he calls *probabilistic modularity*, or **(PM)**. In this section, I will argue that because **(PM)** licenses the same range of causal inferences as **(SAC)**, **(PM)** is a modularity principle of the kind we are looking for. Moreover, that it can act as a bridge principle linking probabilistic dependencies with graphical dependences, **(PM)** has the right features to serve as a modularity principle for the manipulated mechanism. I will also argue, however, that **(PM)** contains much more than a modularity principle, and in the next chapter that this addition is extra baggage that makes **(PM)** an easy target for critics of modularity.

Woodward's **(PM)** makes explicit how to represent the results of interventions in causal graphs. Specifically, **(PM)** claims that, in a correct causal model, the parents (direct causes) of a variable (or set of variables) will screen that variable from interventions into any other variable or set of variables.

**(PM)** Suppose a set of variables $\mathbf{V}$, and $Z, Y \in \mathbf{V}$, and $Z$ is distinct from $Y$. Then

$$\forall Y \forall Z P(Y \,|\, \text{parents}(Y)) = P(Y \,|\, \text{parents}(Y) \& \,\text{set}(Z))$$

(Woodward, 2003, p. 340).

**(PM)** captures the idea that given a variable $B$, and the set of $B$'s direct causes (parents($B$)), interventions into some distinct variable $A$ won't disrupt those causal relationships—they will survive an intervention into $A$, even when $A$ is among $B$'s di-

rect causes. To give a dramatic if overly simple example, whatever causal relationship exists between the location of the moon and the changing of the tides, my experimental intervention into a squid giant axon preparation will not alter that causal relationship. (And if, somehow, magically, they did, wouldn't we wonder if there was a causal connection linking the two after all?)

Given that a dependency (the probabilistic analog to a statistical correlation) between $A$ and $B$ could arise when $A$ causes $B$, and given that interventions represented by the set operator cut the causes of the intervened into variables, then **(PM)** tells us that whatever variables correlate with the intervention are effects of that intervention. Thus, it appears that **(PM)** justifies drawing causal conclusions from the dependencies that arise during an experiment, and hence justifies the same range of causal inferences as does **(SAC)**. I turn now to argue this point.

## The Range of Causal Inference Licensed by (PM)

**(PM)** can be analyzed, as I will show, into two distinct principles. These principles are graph-theoretic analogs to **(SAC)** and **(ACC)**. Because **(SAC)** is a modularity principle, its analog in **(PM)** is a modularity principle too, because it licenses the same kinds of inference. **(ACC)** is not a modularity principle, so **(PM)**, *qua* modularity principle, is carrying around some extra baggage. The significance of this extra baggage will become apparent in the next chapter, when I show that some attacks on **(PM)**, *qua* modularity principle, find purchase only in the analog of **(ACC)**, which is not a modularity principle. For now, the important point is that **(PM)** is in fact a modularity principle, and as a bridge principle linking experimental manipulation with causal graphs, is a candidate formulation of modularity that we can use as a starting point for defending modularity as a general inferential principle.

**(PM)**, as formulated above, claims that interventions into some variable $B$ will

be independent of every other distinct variable $A$, conditional on $A$'s direct causes. This way of putting it suggests four different cases for evaluating **(PM)**: When $B$ is a direct cause of $A$ (when $B$ is a member of parents($A$)), when $B$ is an indirect cause of $A$ (when $B$ is a cause but not in parents($A$), when $B$ is an effect of $A$, and when $B$ is neither a cause nor an effect of $A$.

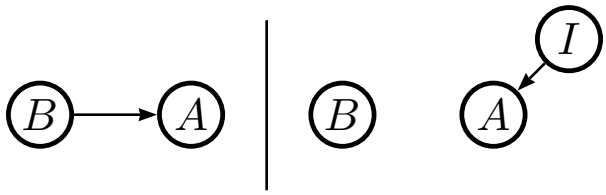


Figure 5.3: $B$ is a cause of $A$, but an intervention into $A$ breaks the arrow leading from $B$ into $A$, representing the idea that interventions should render the intervened-into variable independent of its non-effects.

**Case 1** *$B$ is a cause of $A$.* **(PM)** claims that $B$'s parents screen $B$ from set($A$). Interventions represented by the set operator are what Woodward calls *arrow-breaking*, and what Eberhardt & Scheines (2007) call *structural*. Such interventions into $A$ cut $A$ off from its causes, as in Figure 5.3. So, where $B$ *was* a cause of $A$, set($A$) has no causes in the model (including $B$). So, if $B$ depends on $A$, such a dependance must be coincidental (since neither can $B$ and set($A$) share a common cause). But, since conditioning on parents($B$) could not render $B$ independent of $A$, there is no reason to think that parents($B$) should screen $B$ from set($A$). How could cutting off the causes of $A$ suddenly render $B$'s parents relevant? But **(PM)** claims that $B$ and set($A$) *are* independent given $B$'s parents. So, if this independence relation holds, it must be because $B$ is independent *unconditionally*[9] of set($A$). Thus, for this case, **(PM)** is a stipulation that an intervention into $A$, in virtue of cutting $A$ from its

[9]See §1.1 for a discussion of dependence, conditional and unconditional.

causes, renders $A$ unconditionally independent of its parents. Thus, **(PM)** entails that $B \perp \text{set}(A)$.

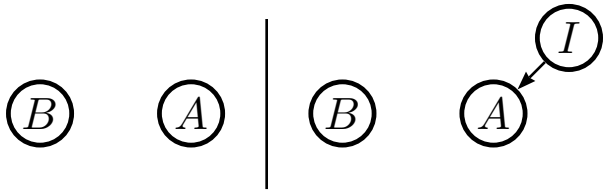


Figure 5.4: Neither $Y$ nor $A$ cause each other.

**Case 2** *Neither $B$ nor $A$ is a cause of the other.* This case turns out to be no different than Case 1, as illustrated in Figure 5.4. In Case 1, $B$ was a cause of $A$, but intervening into $A$ broke that causal connection under the stipulation that $B$ was not a cause of set($A$). When $B$ is already not a cause of $A$, then $B$ is also not a cause of set($A$)[10], and the same logic applies: $B$'s parents are irrelevant to the independence of $B$ and set($A$), even when $B$ and $A$ share a common cause. Thus, since **(PM)** claims that $B$ is independent of set($A$) conditional on parents($B$), that independence must hold because $B$ is unconditionally independent of set($A$). Thus, again, **(PM)** entails that $B \perp \text{set}(A)$.

If we combine the results of Case 1 and Case 2, we get an interesting result: In neither case is $A$ a cause (direct or indirect) of $B$. Thus, **(PM)** entails that if $A$ does not cause $B$, then $B \perp \text{set}(A)$.

But what when $A$ is a cause of $B$?

**Case 3** *$A$ is a cause of $B$.* $A$ can be a direct cause of $B$, or just an indirect cause of $B$. When $A$ is an indirect cause, in the simplest case, $A$ will be a direct cause of

[10]Note that it is not generally true that $B$ will not be a cause of interventions into $A$; but (by definition, see §1.2), $B$ will not be a cause of ideal interventions into $A$, represented by the 'set' operator.

some other variable that is itself a direct cause of $B$. More complex cases will involve more intermediate variables. So, there are two sub-cases to consider.

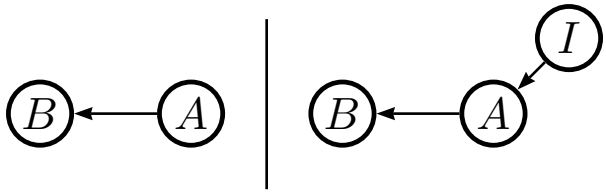


Figure 5.5: $A$ is a direct cause of $B$.

*Case 3.1* When $A$ is a direct cause of $B$, as in Figure 5.5, then so too is set($A$). Since the parents of $B$ contains of all $B$'s direct causes, then set($A$) is in parents($B$). Therefore, parents($B$) = (parents($B$)& set($A$)), and the conditional independence claimed by **(PM)** holds trivially. Thus, **(PM)** entails that $B \perp \text{set}(A) | \text{parents}(B)$.

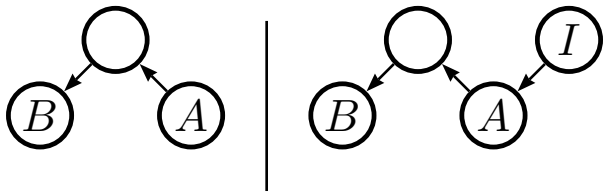


Figure 5.6: $A$ is an indirect cause of $B$.

*Case 3.2* When $A$ is only an indirect cause of $B$, as in Figure 5.6, then **(PM)** claims that $B$'s parents will screen it off from set($A$). **(PM)** entails again that $B \perp \text{set}(A) | \text{parents}(B)$.

Again, we can combine the two cases to get an interesting result, for in both cases is $A$ a cause of $B$, and in both cases **(PM)** entails the same result. **(PM)** entails, therefore, that if $A$ is a cause $B$, then $B \perp \text{set}(A) | \text{parents}(B)$.

We can formalize these two principles comprising **(PM)**.

**(PM)** Suppose a set of variables $\mathbf{V}$, and $Z, Y \in \mathbf{V}$, and $Z$ is distinct from $Y$. Then $\forall Y \forall Z$

**(PMa)** If $Y$ is a non-descendant of $Z$, then $Y \perp \text{set}(Z)$.

**(PMb)** If $Y$ is a descendant of $Z$, $Y \perp \text{set}(Z) | \text{parents}(Y)$.

With this analysis of **(PM)**, we can see clearly that **(PMa)** is a graph-theoretic analog of **(SAC)**. **(SAC)** claims that when $A$ does not cause $B$ and both are correlated, interventions into $A$ will not be correlated with $B$. **(PMa)** claims that when $B$ is not a descendant of $A$, then set($A$) will be probabilistically independent of $B$. The difference is twofold. First, the definition of the set operator carries additional assumptions about what counts as an intervention, whereas **(SAC)** takes intervention as an unanalyzed primitive. Second, **(PMa)**, in making a claim about probabilistic independence and graphical relations, brings the axioms of probability and graph theory along as additional assumptions as well, where again **(SAC)** takes correlation as an unanalyzed primitive concept. Thus, **(PMa)** is a modularity principle, and because it contains **(PMa)**, so too is **(PM)**.

In addition to containing a modularity principle, **(PM)** contains something else: **(PMb)**, which is a graph theoretic analog of **(ACC)** (one that bears a close relation to the Markov principle, as I will show in the next chapter). **(PMb)** claims that when $B$ is a descendant of $A$, then set($A$) will be probabilistically independent of $B$ *conditional on B's parents*. This is an important distinction, because it does not entail that set($A$) will necessarily depend on $B$ (as **(ACC)** claims), but that $B$'s parents will screen it from any intervention into $A$. I will have more to say about this feature of **(PMb)** in the next chapter.

## 5.4 Conclusion

In this chapter, I have aimed to make the assumption of modularity explicit and clear.

I have shown that causal inference from experimental manipulation requires two assumptions, the asymmetry of causal correlation **(ACC)** and the symmetry of accidental correlation, **(SAC)**, and that of these two, **(SAC)** is best viewed as a primitive articulation of the modularity principle.

I have argued, using **(SAC)** as a measuring stick, that many well-known principles that happen to be called "modularity" do not in fact articulate the same intuitions about causal inference that **(SAC)** does. I have also argued that one articulation, Probabilistic Modularity **(PM)**, due to Woodward (2003), does not only articulate the right intuitions, but provides an explicit bridge from experimental observations to the tools of causal modeling, making **(PM)** the best articulation of modularity to use in my broader defense of modularity.

I have also argued, in the course of comparing **(PM)** to **(SAC)**, that we can usefully analyze **(PM)** into two distinct principles, **(PMa)**, and **(PMb)**. The first, **(PMa)**, articulates the same thesis as **(SAC)**, making it a modularity principle, where the second, **(PMb)** does not articulate the intuitions behind modularity; instead, as I will show in the next chapter, it articulates a completely different idea, that causes are Markovian in nature.

In the next chapter, I turn to consider a pair of arguments that criticize **(PM)** on the grounds that it is closely bound up with another controversial principle, the causal Markov condition. I will show that most of this criticism really falls on **(PMb)**—the non-modularity principle half of **(PM)**, and that these attacks present us with good reason for dropping **(PMb)** from our analysis. The remaining attacks fall on **(PMa)**, but, as I will show, rely on an highly questionable assumption.

In the final two chapters of the dissertation, I will present and argue for a principled account of modularity based on **(PMa)**.

# Chapter 6

# Modularity and the Causal Markov Condition

Modularity bears a close relationship to a distinct causal inference principle called the causal Markov condition. Some authors, notably Nancy Cartwright, have argued that there are many mechanisms in biology (and elsewhere) that violate the causal Markov condition. Do these mechanisms also violate modularity, and if so, do they pose a threat to my account of the Manipulated Mechanism? In this chapter, I argue for two claims. The first claim is that my account of the Manipulated Mechanism *can* account for mechanisms that violate the causal Markov condition. The second claim is that Woodward's (2003) articulation of modularity, **(PM)**, although a good first approximation of a formal modularity principle, is too strong as a principle of modularity, and therefore needs to be weakened. These two claims are closely related. Mechanisms that violate the causal Markov condition, I will argue, do not appear to violate the intuitive notion of modularity presented in the previous chapter. But they do sometimes violate **(PM)**, in virtue of violating one of its components, a principle I introduced in the previous chapter called **(PMb)**. This tension suggests that **(PMb)**

is not an expression of some facet of modularity, and does not therefore belong in a formal articulation of modularity. I close the chapter by arguing for the rejection of **(PMb)** as a part of a modularity principle, and for the more general claim that modularity and the causal Markov condition are perhaps not so closely related as has been thought.

Consider the following argument:

1. The Manipulated Mechanism requires Modularity.
2. Modularity implies the causal Markov condition.
3. Many (indeterministic) biological mechanisms violate the Markov condition.
4. Therefore many (indeterministic) biological mechanisms fall outside the scope of the Manipulated Mechanism.

Insofar as biology and neuroscience are littered with indeterministic mechanisms, we should like to avoid the conclusion of this argument. The first premise is straightforwardly true. The second premise has been offered and defended by Hausman & Woodward (1999). Cartwright (1999a) has argued for the plausibility of the third. The conclusion follows by repeated application of *modus tolens*. The above argument is obviously valid, and is, as I will argue, sound—for certain readings of the modularity principle, readings that I will argue are too strong. In this chapter, I will argue that the conclusion can in fact be avoided, because the first premise is only true under an overly strict reading of modularity, and because we can reformulate a weaker

articulation of modularity that is less strict, rendering the third line false—and the argument, therefore, unsound.

But first, consider a biological example. Patients infected with HIV typically, after a period of latency, develop acquired immunodeficiency syndrome (AIDS), which is characterized by a wide variety of symptoms resulting from opportunistic infections. Within any given period in the typical range of clinical latency (typically between several months and twenty years), it is a matter of probability as to whether a particular HIV-positive patient will develop AIDS. The HIV virus is the common cause of the symptoms compresent in the syndrome, although the production of the symptoms occurs through several paths, including the infection of $CD4^+$ T-cells, macrophages, and microglial cells. Nevertheless, because of the timecourse of AIDS, the symptoms are all strongly correlated without being a cause of each other, such that having one symptom is often a greater predictor of having other symptoms than is being HIV-positive. Moreover, there is no single factor, not even T-cells, that we could point to as a single process responsible for all of the symptoms, because HIV works through many channels simultaneously (even though most symptoms do develop as a result of a low T-cell count). (Wikipedia, 2010)

The interesting structural feature to this example is that there is a common cause—HIV infection—with many effects—the various symptoms. Each effect has a certain probability of appearing, and knowing that a patient is HIV-positive increases the probability that any particular symptom is present. But the probability that one particular symptom is present is increased even more when we know that a patient has one or more *other* characteristic symptoms. Put slightly differently, if we know that someone is HIV+, finding out that they have one symptom makes it even more likely that they have a second, beyond knowing only that they were HIV+.[1]

[1]Thanks to Frederick Eberhardt for putting the point in this way.

Thus:

$$S_1 \leftarrow \quad H^+ \quad \rightarrow S_2$$

but

$$Pr(S_1|H^+) \quad > \quad P(S_1)$$

$$(1)Pr(S_1|S_2\&H^+) \quad > \quad P(S_1|H^+)$$

This characterization of AIDS represents a failure of the screening off principle. A variable (or set of variables) $C$ screens one variable $A$ from another $B$ when, given the state of $C$, finding out the state of $B$ does not increase the probability of $A$ occurring. Put a little differently, $B$ adds no information about the probable state of $A$ beyond what $C$ already gives us. Recall from Chapter 1 that the causal Markov condition entails that common causes screen their children from each other, that is, knowing the common cause tells us everything there is to know about either of the effects, and that learning something about one of the effects will not tell us anything new about the other one. When $Z$ is a sibling of $X$ (and is not also an effect of $X$), then the parents of $X$ (which are, in this case, also the parents of $Z$) will screen $X$ from $Z$. Therefore, from **(CM)**:

$$Pr(X|\,\text{parents}(X)) = Pr(X|\,\text{parents}(X)\&Z)$$

But in the case of AIDS (from (1) above):

$$Pr(S_1|\,\text{parents}(S_1)) < Pr(S_1|\,\text{parents}(S_1)\&S_2)$$

The common-cause in this example, being HIV positive, fails to screen its effects from each other, because the presence of one symptom is a better indicator of the presence of other symptoms than is knowing a patient is HIV positive. The reason is simple: Patients that are HIV positive may not yet—or ever—exhibit any symptoms of AIDS. But once one symptom develops, many others follow very quickly. So,

although AIDS is a cause of the symptoms, knowing whether a patient has a particular symptom tells us more about whether that patient has any of the others.[2] Thus, AIDS violates the Markov condition. If my argument above is correct, then this is a serious problem for my account of the Manipulated Mechanism, if it cannot account for the mechanism for AIDS or other syndromes.

In this chapter, I consider the consequences for the modularity principle of failures of the causal Markov condition—exhibited, for example, by syndromes. In §6.1, I present Hausman and Woodward's argument that **(PM)**, a formulation of the modularity principle, entails **(CM)**, a formulation of the causal Markov condition. I also present a reason to suspect the argument's conclusion, namely that Hausman and Woodward have misconstrued what it means to be an intervention in an indeterministic system.

Then, I turn to consider two different kinds of causal structure that give life to this suspicion, causal structures that violate **(CM)** without violating **(PM)**. In §6.2, I consider a class of examples similar to the AIDS example above offered by Cartwright (1999a). These examples, which I call Polluting Factories, have a common cause structure in which two effects are more closely correlated with each other than with their common cause. I argue that Polluting Factories, despite violating **(CM)**, do not necessarily violate **(PM)**.

In §6.3, I consider a different class of examples, which I call Degenerate Chains, in which a remote, indirect effect appears to exert a causal influence across temporal or physical gaps. For example, the 1989 Exxon *Valdez* oil spill is considered to have had a direct influence on the 1993 collapse of the herring populations in Prince Henry Sound, despite a large temporal and causal gap between the two events. I argue that

[2]Moreover, whether there is a single intermediate cause between the HIV infection and the appearance of each symptom that could explain this correlation is an empirical question; even if we might think such an intermediate cause was likely, there is no *a priori* reason why there *must* be one.

Degenerate Chains, like Polluting Factories, do not necessarily violate **(PM)**.

On the basis of these two examples, I draw two conclusions. First, I argue that **(PM)** does not entail **(CM)** because Hausman and Woodward's reasoning rests on a conceptual confusion about the relationship between indeterministic variation and interventions. Second, I argue that, contrary to what we might conclude on the basis of **(PM)**, the contexts in which Polluting Factories and Degenerate Chains *satisfy* **(PM)** are contexts that, ironically, seem to contradict the *spirit* of **(PM)**, that interventions should not disrupt the probability distribution within a model.

In §6.4 I resolve this tension by arguing that **(PM)**, as a formulation of the principle of modularity, is too strong, I draw upon my analysis of **(PM)** into **(PMa)** and **(PMb)** from the previous chapter. I will argue, on the basis of these examples, that **(PMb)** is not a proper part of the best analysis of modularity, and that we should instead advert to **(PMa)** alone as our modularity principle. I will show that **(PMb)** is a heavily conditionalized version of the causal Markov condition. But it has the odd feature that, given some mechanism like a Polluting Factory or a Degenerate Chain that violates the causal Markov condition, that interventions into certain parts of those mechanisms should *alter* the mechanism to satisfy the Markov condition. Yet, such an alteration seems to extend beyond the limited reach that interventions should have, and indeed seems a violation of the very spirit of **(PMb)**. I close the chapter by concluding that a **(PMb)** is an ancillary principle, and not a part of modularity itself.. I show that such a move does, unfortunately, increase the ambiguity of certain kinds of causal inference, but does not yield false inferences about Polluting Factories, as **(PM)** would. I also argue that the weakened condition is more consistent with the spirit of the modularity condition.

In the next chapter, I turn to consider a class of arguments against modularity to which **(PMa)** remains vulnerable: That modularity requires (or just is the require-

ment) that the various components in a mechanism be independently manipulable.

## 6.1 An Argument that (PM) Entails (CM)

Let us begin by considering the first premise in the argument that opened this chapter:

1. Modularity entails the causal Markov condition.

The causal Markov condition, recall, connects causal graphs with probabilities—it is a principle for deriving causal structure from observed correlations (see §1.1). The Markov condition (notice that I have not said 'the *causal* Markov condition') states that, in a directed graph, each variable (node) is *independent* of its non-descendants conditional on its parents. Somewhat informally: Once we know the states of a variable $X$'s parents, learning the state of any other variable (except those that are descendants of $X$, and of course $X$ itself) adds nothing to our knowledge of $X$. One way to express this independence is to write:

**Markov Condition** Where $Y$ is not a descendant of $X$,

$$Pr(X\&Y\,|\,\text{parents}(X)) = Pr(X\,|\,\text{parents}(X)) \cdot Pr(Y\,|\,\text{parents}(X)).$$

The Markov condition becomes the *causal* Markov condition when we interpret the kinship relationships as causal relationships, *i.e.*, 'parent' = 'direct cause', 'ancestor' = 'direct or indirect cause', *&c.*

Hausman & Woodward (1999) have argued for the claim that modularity entails the causal Markov condition. Roughly speaking, their argument is that, because interventions should not disrupt the causal structure of a system, then what is true during an intervention is (ideally) true outside of an intervention as well. Modularity places constraints on the causal structure during an intervention that are related to

the constraints placed on causal structures by the causal Markov condition; then, outside of any intervention, a modular system should therefore conform to the causal Markov condition (should be, in the parlance, "Markovian").

Here is their argument in detail. They begin by supposing we have some set of variables $\mathbf{V}$ that we use to represent a causal structure, and offer a characterization of what it means to intervene into a variable in this set. "[T]he relevant notion of an intervention that sets the value of [any variable in $\mathbf{V}$] $X$ is," they tell us, "(roughly) just that of something that is a direct cause of $X$ and that bears no causal relations to the other variables under consideration except those that arise from its directly causing $X$" (p. 535). Moreover, the intervention must be exogenous (must not have a cause within the set of variables $\mathbf{V}$), and it must be uncorrelated of every other variable that is not a child of $X$. (p. 535–536)

Hausman & Woodward assume the following three principles (detailed below): causal sufficiency; a principle that they call **(CM1)**[3]; and the principle that interventions can be treated as independent random variables (p. 573).

*Causal sufficiency* is the claim that we have not left out of $\mathbf{V}$ any relevant common causes (which could, if otherwise left out of $\mathbf{V}$, create spurious correlations among some variables.) (p. 527).

**(CM1)** is the claim that probabilistic dependence among variables must result from a causal dependence among them:

**(CM1)** If $X$ and $Y$ are probabilistically dependent, then either $X$ causes $Y$ or $Y$ causes $X$ or $X$ and $Y$ are effects of some common cause $Z$ in the set of variables $\mathbf{V}$ (p. 524).

[3]Hausman and Woodward use 'CM' to name principles related to the causal Markov condition, and 'PM' to name principles related to modularity.

Worth noting at this point is that **(CM1)** and causal sufficiency together entail that all of the exogenous variables in $\mathbf{V}$ are independent of each other.

The principle that interventions can be treated as independent random variables states simply that we can represent an intervention into $X$ by replacing occurrences of $X$ in our model with a random variable set $X$. Moreover, interventions into $X$ render it independent of its parents (and indeed all of its non-descendants); so the claims that set $X$ is an *independent* random variable means that set $X$ is unconditionally independent of $X$'s descendants.

Hausman & Woodward then use the assumption that interventions can be treated as independent random variables to construct a formal notion of modularity, which they call *Probabilistic Modularity*, or just **(PM)**.

**(PM)** When $Z$ is any set of variables distinct from $Y$, and the values to which the variables in $Z$ are set lie within the relevant range, then

$$P(Y|\,\text{parents}(Y)) = P(Y|\,\text{parents}(Y)\&\,\text{set}\,Z)$$

(p. 573). Equivalently:

$$Y \perp \text{set}\,Z|\,\text{parents}(Y).$$

The intuition driving this formulation is simply that if we could not intervene into $Z$ independently of $Y$, then the probability distribution over $Y$ conditional on $Y$'s parents would change when we intervened into $Z$. When we can intervene into $Z$ independently of any other variables, then **(PM)** holds for all $Y$. When we can intervene into any variables independently of any other variable, then **(PM)** holds also for all $Z$.

Enough premises: To the argument itself, then. Take any two distinct variables $X$ and $Y$ in $\mathbf{V}$; **(PM)** implies that $Y$ is independent of interventions on $X$, given $Y$'s

parents (the direct causes of $Y$) represented in $\mathbf{V}$:

$$P(Y|\,\text{parents}(Y)) = P(Y|\,\text{parents}(Y)\&\,\text{set}\,X) \tag{6.1}$$

Now, if, in addition, $X$ does not (directly or indirectly) cause $Y$, **(PM)** implies that each ancestor of $Y$, $Z$, is also independent of interventions on $X$ conditional on its ($Z$'s) parents.

$$P(Z|\,\text{parents}(Z)) = P(Z|\,\text{parents}(Z)\&\,\text{set}\,X) \tag{6.2}$$

In other words, interventions on $X$ render it independent of each and every of $X$'s non-descendants.

Both (6.1) and (6.2) together imply that interventions into $X$ are *unconditionally* independent of $Y$. This holds for the following reason. If $X$ were a cause of $Y$, then we would expect that $Y$ *would* be dependent on interventions on $X$; **(PM)** implies that $Y$'s causes will screen $Y$ off from interventions on $X$ in this case. But when $X$ is not among $Y$'s ancestors, there are no variables 'causally between' $X$ and $Y$. Thus, there is nothing to screen off: interventions on $X$ must be independent from $Y$ unconditionally. Hausman & Woodward call this result **(MOD*)**.

**(MOD*)** For all distinct $X$ and $Y$ in $\mathbf{V}$, if $X$ does not cause $Y$, then

$$P(Y\&\,\text{set}\,X) = P(Y)\cdot P(\text{set}\,X)$$

(p. 576).

**(MOD*)** claims that interventions into $X$ render $X$ independent of every other variable, save $X$'s descendants. So far, so good.

Now, Hausman & Woodward ask us to back up for a moment, and consider what happens when we hold parents($X$) constant. If $X$ is indeterministic, $X$ can vary even when we hold parents($X$) fixed. "So in the circumstances in which parents($X$) is

unchanging," Hausman & Woodward say, "either $X$ varies spontaneously or because of causes that have no causal relation to any other variables except in virtue of causing $X$" (p. 576). Since we are holding parents($X$) fixed, these additional causes of $X$ must not appear in $\mathbf{V}$. At any rate, **(CM1)** and Causal Sufficiency ensure that each of these sources of variation is unconditionally independent of every other exogenous variable, and therefore count as interventions into $X$. That is, **(CM1)** and Causal Sufficiency ensure that there is no source of indeterministic variation in $X$ that does not also meet the criteria for an intervention into $X$.

Thus, and this is the crucial turn in the argument, "since changes in $X$ conditional on parents($X$) count as interventions with respect to $Y$, changes in $X$ conditional on parents($X$) must be independent of the same things that $set - X$ is independent of" (p. 576). Therefore, the following bi-conditional holds, which I call *Modular Markov*, or simply **(MM)**[4]:

**(MM)**

$$P(Y \,\&\, \text{set}\, X) = P(Y) \cdot P(\text{set}\, X)$$

$$\Leftrightarrow$$

$$P(X \,|\, Y \,\&\, \text{parents}(X)) = P(X \mid \text{parents}(X)).$$

Notice that the left-hand side of **(MM)** is **(MOD*)**; Hausman & Woodward call the principle on the right-hand side **(CM2)**. Because Hausman & Woodward have already argued for **(MOD*)**, they conclude that **(CM2)** must therefore also hold.

Their final move is to notice that the Markov condition is simply the conjunction of **(CM1)** and **(CM2)**. Since Hausman and Woodward have assumed **(CM1)**, and

[4]Hausman & Woodward call this bi-conditional '**(=)**', which seems confusing. I have altered their nomenclature for clarity.

have derived **(CM2)** from **(PM)**, then the conjunction of **(CM1)** and **(PM)** implies the Markov condition. QED.

Thus have Hausman & Woodward argued from modularity in the guise of **(PM)** (plus auxiliary assumptions) to the causal Markov condition; moreover, this argument is designed to hold in indeterministic contexts (they provide a separate argument for deterministic contexts that closely parallels this one).

## A Weakness in the Argument

I first should note that other authors have raised problems for this argument. Steel (2006) has argued that **(CM1)** is doing most of the heavy lifting in the argument, and that **(PM)** is in fact superfluous. Cartwright (2002) has argued that both **(PM)** and **(CM2)** are separately derivable from the background assumptions, and hence it is a mistake to think that one therefore entails the other.

I raise a different issue for the argument. I am willing to entertain that **(PM)** plays a central role in the argument, and that the background assumptions are not doing any kind of inferential heavy lifting. Instead, I find Hausman and Woodward's justification for the bi-conditional **(MM)** suspect. The justification is meant to show how two apparently distinct claims link up: **(MOD*)**—a claim about causal systems under intervention—and **(CM2)**—a claim about causal systems not under intervention. The justification centers around the claim that indeterministic variation in a variable counts as an intervention. Hausman and Woodward explain:

> By assumption, in addition to its represented direct causes (parents($X$)), it [$X$] has some unrepresented causes whose effect is summarized by the error term $U_X$ in the equation for $X$. By definition, these are direct causes, and, given causal sufficiency, they bear no causal relationship to any vari-

> able other than $X$ apart from those which result from their being direct causes of $X$. $U_X$ thus satisfies the definition of an intervention with respect to the [sic] $X$... (p. 553)

We might wonder why the indeterministic variation in $X$ (represented by Hausman and Woodward by 'an error term $U_X$') will always satisfy the definition of an intervention with respect to $X$. An intervention is not simply a cause of some variable $X$, but has close restrictions placed on it: 'The crucial point', Hausman and Woodward stress, 'is that an intervention with respect to $X$ is a direct cause of $X$ that has no causal relations to any of the variables in $\mathbf{V}$ [the model] except in virtue of being a direct cause of $X$' (p. 536). Elsewhere (p. 553), Hausman and Woodward make the further claim that **(CM1)**, in virtue of these causal restrictions, entails that interventions on $X$ cannot be correlated with any other variable in the model (except for $X$'s effects, of course).

And yet, as I will now turn to show, it is conceivable for a causal system to violate the causal Markov condition without violating **(MOD*)** or **(CM1)** (during an intervention), which if true demonstrates that the two halves of **(MM)** can come apart. The difficulty is this. Suppose that we have a causal system in which one indeterministic variable $X$ is causally unrelated to another $Y$, and yet the indeterministic variation in $X$ is nevertheless correlated with the variation in $Y$. Such a system violates **(CM1)**. Yet, in such a system, it is conceivable that there is some intervention on $X$—an intervention that is *not* simply the indeterministic variation in $X$, but rather one that is capable of overriding this variation—that renders $X$ independent of $Y$. Such a system under intervention does not violate **(CM1)**. Thus, one and the same system can violate **(CM1)** and satisfy **(CM1)** depending on whether it is being intervened into. This contradicts Hausman and Woodward's justification for **(MM)**.

In the following sections, I present two different kinds of causal structure with

this kind of feature, wherein the indeterministic variation in one or more variables violates **(CM)**, yet where an external intervention can restore it. The possibility of such systems suggests that the link between external interventions and indeterministic variation is not nearly so close as Hausman and Woodward believe.

## 6.2 Polluting Factories

In this section, I want to consider one way in which the causal Markov condition can be violated: By common cause structures in which the joint effects are more closely correlated with each other than they are with their common cause. Call these, for reasons that will become clear below, *Polluting Factories*. In this section, I begin by describing how such a structure violates Markov. I then turn to Nancy Cartwright's arguments that the causal Markov condition is not generally satisfied by probabilistic causal structures because Polluting Factories are strikingly common. Although Cartwright uses her discussion as an argument for causal pluralism (and, by extension, as an argument against any domain-general system of causal inference), I put her discussion to a slightly different use: To show one central point of departure between the causal Markov condition and modularity. I will argue that that common cause structures that violate the Markov condition do not necessarily lead to violations of modularity, and hence that we have good reason to think that modularity does *not* entail the causal Markov condition, *contra* Hausman and Woodward.

Cartwright (1999b) observes that the causal Markov condition encompasses two different constraints. One constraint is a proscription against causal influence across temporal gaps (*i.e.* Degenerate Chains, which I address in §6.3 below). The second constraint, and the one that Cartwright (1999b) is interested in, ensures that the parents of a variable *screen* that variable *off* from its siblings. A variable (or set

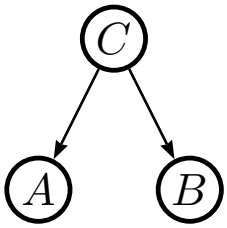


Figure 6.1: $C$ is a common cause of $A$ and $B$

of variables) $C$ screens one variable $A$ from another $B$ when, given the state of $C$, finding out the state of $B$ does not increase the probability of $A$ occurring, that is. Put a little differently, $B$ adds no information about the probable state of $A$ beyond what $C$ already gives us.

Now, consider a common cause structure as in Figure 6.1. Since $C$ is the only cause of $A$ and of $B$, and neither $A$ nor $B$ cause each other, the causal Markov condition (as I mentioned above) implies that $C$ screens $A$ off from $B$: Given knowledge of the state of $C$, knowing the state of $A$ adds nothing to our knowledge of $B$ and *vice versa*. Thus, $A$ and $B$ are conditionally independent given $C$:

$$Pr(A\&B|C) = Pr(A|C) \cdot Pr(B|C)$$

This conditional independence represents the cross-constraints placed on the probability distribution over $A$ and $B$ as a result of the underlying causal structure, namely $A$ and $B$'s sharing a common cause $C$. And, centrally for the causal inference project of Spirtes, Glymour, & Scheines (1993), this constraint can be used as a marker to identify a common cause (or at least that there is a common cause) or an intermediate cause from observational data about $A$ and $B$—namely, that when $A$ and $B$ are observed to be dependent, yet conditionally independent on $C$, we can conclude that $C$ is a common cause of $A$ and $B$.

Cartwright offers a concrete example to illustrate why the Markov condition works

to help us identify common causes. As it happens, candy consumption is strongly (albeit negatively) correlated with divorce rates. But there doesn't seem to be a good reason to posit a causal relation between candy consumption and divorce rates—How could (in general) giving up candy break up a family? Why should discovering infidelity lead to a loss of appetite for sweets? But, if we additionally consider the effects of aging, the solution becomes clear: It is unsurprising to learn that children have a stronger preference for candy than do adults, nor is it unsurprising to learn that children typically don't divorce, since they cannot marry in the first place. Once we condition on the age of each subject, we see that age neatly accounts for both effects: Children have a stronger preference for candy and are less likely to divorce (since they are too young to marry), while adults tend to replace their taste for candy with a taste for Brussels sprouts, and are more likely to divorce (since they have surpassed the legal age to marry).

Thus, once we know that a person is a child, knowing that they are not divorced does not tell us anything new—we could already infer a preference for candy from their age. Likewise, once we know that a person is an adult, knowing their preference for candy is uninformative about the chances of their being divorced. The correlation between candy and divorce goes away, because age screens candy consumption from divorce rates. The screening-off condition of the causal Markov condition entitles us to conclude that age is a common cause of candy consumption and divorce rates.

Crucial to this example, Cartwright notes, is that the causal mechanisms by which age diminishes taste for candy and by which age increases the likelihood of divorce are quite distinct. We can explain how age screens off candy consumption from divorce by telling two different stories. With candy, we might tell a neuro-physiological story about changing metabolic needs, the development of will-power, and the effects these changes have on the eating behaviors of young adults. But with divorce, we might

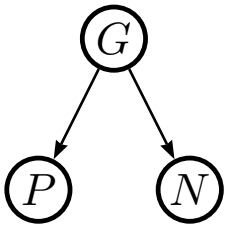


Figure 6.2: In glycolosis, glucose ($G$) is converted into a product/by-product pair, pyruvate ($P$) and NADH ($N$).

tell a sociological-legal story about the contractual institution of marriage and its dissolution, and in particular about how we deny children the right to enter into legally binding contracts. So we should not be surprised that age, here treated as a very coarsely-grained variable, renders candy consumption and divorce rates independent, because the mechanism by which age has an effect on candy consumption and the mechanism by which age has an effect on divorce rates are themselves so very different. Thus, Cartwright notes that the Markov condition will correctly identify common causes when the joint effects are brought about by distinct processes, as with candy consumption and divorce rates.

But why, Cartwright wonders, should we think that common causes *invariably* produce their effects via completely distinct processes? Many processes result in a product, and simultaneously with that product, one or more by-products. These by-products are not produced via a (completely) independent process from the products themselves. Glycolosis, to take a biological example, is a process of converting glucose into pyruvate which also produces NADH as a by-product. We might represent this metabolic process using a causal graph as in Figure 6.2.

Glycolosis does not proceed by first making some glucose into pyruvate, and then making other glucose into NADH via completely different sequences. Rather, NADH molecules are the tailings of pyruvate production: the two are products of the same

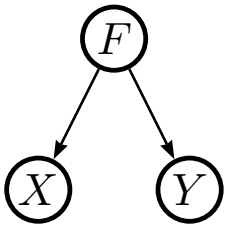


Figure 6.3: Factory $F$ produces chemical $X$, and pollutant $Y$ as a by-product.

process.

Cartwright worries that in indeterministic cases, where the action of the process is chancy, and does not completely determine whether product and by-product are produced (as often happens in molecular pathways), this kind of example will run afoul of the Markov condition.[5] It will run afoul of Markov when the following three conditions are met: There is a common cause that produces a product/byproduct pair via a single process as in the example of glycolosis above; The common cause's working is chancy, *i.e.* it doesn't always result in a product; And when the common cause *does* work, the product and by-product are invariably both produced together.

As an illustration, Cartwright constructs an example she calls the Polluting Factory (see Figure 6.3). The city's sewage plant requires chemical $X$ to treat raw sewage. They purchase chemical $X$ from a nearby factory $F$. Unfortunately, factory $F$ only successfully produces chemical $X$ about 80% of the days it is operational. Moreover, when it does produce chemical $X$, it also produces nasty pollutant $Y$ as a necessary by-product.

In this example, the factory ($F$) is a common cause of both chemical $X$ and pollutant $Y$. Since $F$ is the one and only cause of both $X$ and $Y$, the Markov condition

[5]In deterministic causation, causes necessitate their effects; thus, common causes necessitate their joint effects, and so $Pr(E_1|C) = Pr(E_2|C) = 1$. This equality ensures that **(CM)** is trivially satisfied: $Pr(E_1|C) = Pr(E_1|C\&E_2) = 1$.

tells us (see above) that $X$ and $Y$ ought to be independent of their common cause $F$:

$$Pr(Y\&X|F) = Pr(X|F) \cdot Pr(Y|F).$$

But notice that in the narrative of the Polluting Factory, the independence does not hold. The probability that $X$ is produced, given that the factory is operating, is 80%; likewise for the byproduct $Y$. Multiplying these together yields:

$$Pr(X|F) \cdot Pr(Y|F) = 0.8 \cdot 0.8 = 0.64$$

But because pollutant $Y$ is a byproduct of the process for making chemical $X$, the probability of $X$ *and* $Y$, given that the factory is operating, is also 80%:

$$Pr(Y\&X|F) = 0.8$$

The screening-off condition, as I've just said, requires these quantities to be the same. But they are not. There is a dependency between $X$ and $Y$ that is not fully accounted for by looking at the common cause $F$; $F$ fails to screen $X$ off from $Y$. So the Polluting Factory violates the screening-off condition, and hence the Markov condition.

One way to make clear how the Polluting Factory violates the Markov condition is to note that the Markov condition encapsulates the intuition that any dependency must have a causal explanation: Either one correlate causes the other, or they share a common cause. But chemical $X$ and pollutant $Y$ share a non-causal correlation; although the common cause $F$ does account for *some* of the dependence between $X$ and $Y$, it does not account for *all* of it. And since there is no further causal connection between $X$ and $Y$, the Markov condition is not satisfied by this structure.

We could, however, take the standard tack of claiming that such a violation of Markov is evidence that there must be some latent variable that should be added to our model, a variable that will account for this stray dependence. We might speculate that both chemical $X$ and pollutant $Y$ are in fact not a direct result of the factory's

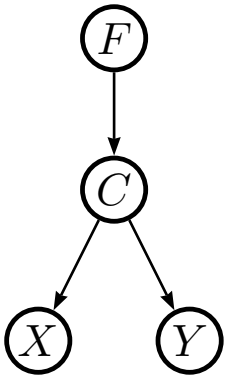


Figure 6.4: Factory $F$ produces chemical $X$ and pollutant $Y$ by way of chemical process $C$.

chancy operating $F$, but by a hidden cause that links the factory operations to a particular chemical reaction $C$ that yields both $X$ and $Y$ and that is itself not chancy. The resulting model is presented in Figure 6.4. On this addition, $C$ is produced by $F$ 80% of the time; and $C$ necessitates both $X$ and $Y$. In this way, we can restore the Markov condition, because

$$\begin{aligned} Pr(X|C) &= 1 \\ Pr(Y|C) &= 1 \\ Pr(X\&Y|C) &= 1 \\ Pr(X\&Y|C) &= Pr(X|C) \cdot Pr(Y|C). \end{aligned}$$

On this view, the Polluting Factory, as presented, is not a plausible counter-example because it leaves out a crucial detail, namely something about how chemical reaction $C$ is the true common cause of $X$ and $Y$. But we should pause for a moment and ask, as Cartwright asks, why we should be willing to commit ourselves to this new story. It seems an empirical question, and not an *a priori* one, whether $C$ really exists and plays this purported role. So it is unfair to insist that there must be some entity to play the role of $C$, *a priori*, as the Markov condition seems to do. Indeed, Cartwright

asks, is the Polluting Factory really so implausible? Regardless of our gut instincts, let us follow her lead for a moment, and suppose that it is plausible, without the further decoration with latent variables.

Cartwright contends that the plausibility of the Polluting Factory rests on two independently plausible claims. The first claim is that there can be genuine macro-level cases of indeterministic causation—that we should not be forced to advert to finer-grained analyses (such as the one above involving $C$) simply because of a need to explain away apparent violations of **(CM)**. The second claim is that joint effects of a common cause can be produced by non-independent processes. The first claim seems reasonable enough, and I won't devote attention to it; at any rate, I certainly do not wish to dispute that point. However, we might think that the second claim, upon which the Polluting Factory rests, is contentious. Cartwright provides two independent reasons to accept the non-independence claim.

Taking the the possibility of genuine macro-level indeterminacy in more detail, consider the range of possible outcomes of the Polluting Factory. When $F$ occurs (when the factory has been asked to produce chemical $X$), the possible outcomes are $(+x, +y)$, $(-x, +y)$, $(+x, -y)$, and $(-x, -x)$[6]. When we allow that causality can be probabilistic, Nature assigns[7] a joint probability over the space of possibilities by assigning a probability to each outcome, $Pr(+x, +x)$, $Pr(-x, +y)$, $Pr(+x, -y)$, and $Pr(-x, -y)$. How are these assignments made? The axiomata of probability place certain constraints: Each probability must be greater than zero, and the sum of the probabilities of all outcomes must come to one. But these constraints are largely notational, reflecting a constructed system for calculating probabilities. The

[6]Where $+x$ indicates the production of chemical $X$, and $-x$ indicates the non-production of chemical $X$, etc.

[7]Cartwright is all too happy to anthropomorphize the external world. I indulge her here just a bit, simply because not much hangs on it. Rather, she seems to find talking about Nature and Nature's actions as a useful shorthand for much wordier and vexed locutions about laws of nature or the metaphysical structure of the universe.

Markov condition is different, Cartwright observes, because it places an ontological (rather than an epistemological) constraint on Nature's assignment of probabilities such that $Pr(+x, +y) \cdot Pr(-x, -y) = Pr(-x, +y) \cdot Pr(+x, -y)$. But "nothing in the concept of causality, or of indeterministic causality," Cartwright claims, "constrains how Nature must proceed." If Nature *was* so constrained, as must be the case for the Markov condition to hold universally, then the Markov condition comes with an implicit—and rather mysterious—metaphysical commitment about the structure of the universe, such that there is this constraint in the world (and not just in our system of representations) that bars Nature from assigning the probabilities as she sees fit. The commitment is mysterious because it simply isn't clear what the world must be like for such a constraint to hold; nor are the proponents of the Markov condition offering any candidate metaphysical conditions—it is outside their job description as epistemologists, as it were.[8]

Cartwright's second consideration is that the existence of Polluting Factories is an empirical matter, not a conceptual matter, and we do seem to be surrounded by them. Syndromes like AIDS, and chemical processes with by-products like glycolosis appear to be real-world examples of Polluting Factories. Bell's elaboration on the EPR experiment is very much a real-world example of a Polluting Factory (albeit one confined to the depths of quantum mechanics)(Bell, 1964; Aspect, Grangier, & Roger, 1981). At any rate, there is no *a priori* reason to think that such cases cannot exist, unless we take the Markov condition as an *a priori* claim (which we do not). Thus, a claim that Polluting Factories cannot exist is an empirical claim, and the onus falls on the claimant for proof thereof.

[8]Indeed, as Craver (personal communications) points out, Clark Glymour and his cohort are insistent that they are not in the business of doing metaphysics, except (as Eberhardt points out to me) insofar as they are willing to reject as murky any metaphysics that does not comport with the Markov condition. In any case, we might wonder that Cartwright sees the causal Markov condition as saddling supposedly metaphysics-neutral causal inference with metaphysical assumptions.

## Polluting Factories and (PM)

So, for the purposes of argument, let us take Cartwright's worry about the Markov condition seriously. We are likely surrounded by Polluting Factories. At any rate, there's nothing conceptually inconsistent about the idea of a Polluting Factory. And, more to the point, there seems nothing about the structure of Polluting Factories that precludes giving a mechanistic explanation of their joint effects. And Polluting Factories *do* violate the Markov condition. Now if, as I discussed in the previous section, **(PM)** entails the Markov condition, then by *modus tollens*, Polluting Factories will also violate **(PM)**. The problem is that they do not violate **(PM)**.

Recall from the previous chapter that we can analyze **(PM)** into the conjunction of two claims, **(PMa)** and **(PMb)**:

**(PM)** Suppose a set of variables $\mathbf{V}$, and $Z, Y \in \mathbf{V}$, and $Z$ is distinct from $Y$. Then $\forall Y \forall Z$

**(PMa)** If $Y$ is a non-descendant of $Z$, then $Y \perp \text{set}(Z)$.

**(PMb)** If $Y$ is a descendant of $Z$, $Y \perp \text{set}(Z) | \, \text{parents}(Y)$.

Suppose, moreover, that we were to intervene into the operating of the factory $F$, to make it run. What does **(PM)** entail about the probability distribution during that intervention? $X$ and $Y$ are both descendants of $F$, thus **(PMa)** is trivially satisfied. **(PMb)** is also trivially satisfied for $X$ and $Y$ because $\text{parents}(X) = \{\text{set}(F)\}$ and $\text{parents}(Y) = \{\text{set}(F)\}$. Now, suppose we were to intervene to generate chemical $X$ at the premises independently of the factory's operation. Here, **(PMb)** is trivially satisfied because neither $F$ nor $Y$ are a descendant of $X$. And, where $F$ is a non-descendant of $X$, **(PMa)** should be satisfied, insofar as our intervention into $X$ is surgical and satisfies the constraints on interventions discussed in Chapter 1. But $Y$

is also a non-descendant of $X$, and there is nothing in the structure of Cartwright's story that tells us how to evaluate the probability $Pr(Y|F\&\,\mathrm{set}(X))$.

On the one hand, we might suppose that intervening into $X$ would not change its dependency on $Y$, that is, that there is no way to manufacture chemical $X$ without also producing pollutant $Y$ as a by-product. In this case, **(PMa)** would be here violated. But on the other hand, there is nothing preventing us from supposing that, for whatever reason, intervening into $X$ could break its dependency on $Y$, that intervening to manufacture chemical $X$ could proceed without the production of pollutant $Y$. Perhaps we need only use a different method than is employed by the factory. In this case, **(PMa)** would be satisfied. And, if it is possible for the Polluting Factory to satisfy **(PMa)** and **(PMb)**, then the Polluting Factory stands as a counter-example to Hausman and Woodward's claims that **(PM)** entails **(CM)**.

To draw this final point out in more detail, let us consider a different kind of causal system that violates **(CM)**, a structure I call a *Degenerate Chain*. A close look at Degenerate Chains will shed light on what has gone wrong for Hausman and Woodward's justification for **(MM)**.

## 6.3 Degenerate Chains and (PM)

In addition to Polluting Factories, **(CM)** is also violated by what I call Degenerate Chains. A Degenerate Chain is a causal structure that consists of a causal chain $X \rightarrow Y \rightarrow Z$, (with no arrow directly connecting $X$ to $Z$), and where $Pr(Z|Y) \neq Pr(Z|Y\&X)$: where $Y$ does not screen its effect $Z$ from its cause $X$. In other words, $X$ provides information about $Z$ beyond what $Y$ provides. Such a chain violates the Markov condition, which requires that $Y$ does screen $Z$ from $X$—that $X$ does not provide information about $Z$ beyond what $Y$ provides. As a technical term, I call this

kind of failure of a variable to screen its causes from its effects a degeneracy. The question before us, then, is: Do Degenerate Chains violate **(PM)**?

I will argue in this section for two claims. First, I will argue that Degenerate Chains sometimes violate **(PM)**, and that when they do, it is in virtue of violating **(PMb)**. Second, I will point out that Degenerate Chains have a curious feature: Degenerate Chains that *do not* violate **(PM)** seem, nevertheless, to violate an intuitive notion that **(PMb)** is meant to capture: That interventions should not alter the probability distribution among the intervention's descendants. This tension, I will argue, gives us reason to think that **(PMb)** has no place in a broad articulation of modularity. I will conclude that squaring **(PM)** with our modular intuitions demands that we jettison **(PMb)** from our analysis of modularity, and that therefore neither Degenerate Chains nor Polluting Factories pose a threat to modularity-based accounts of mechanism.

As one might expect, jettisoning **(PMb)** comes at an inferential cost. I close this section by examining what inferential power is lost by abandoning **(PMb)** as part of our analysis of modularity. I show that in certain inferential contexts, **(PMb)** will yield a non-singleton equivalence class of causal structures. Although **(PMb)** will not always yield definitive conclusions, I will argue that the true causal structure is guaranteed to be a member of the equivalence class **(PMb)** yields up. In contrast, **(PM)** will actually yield an incorrect result in these same contexts. I conclude, on this basis, that the inability of **(PMb)** to yield a single, correct causal structure is not in fact a total loss in inferential power, but a gain (in that **(PMb)** alone will never yield incorrect results), and that this provides additional reason to abandon **(PMb)** from our analysis of modularity.

Degenerate Chains come in two varieties: transitive and intransitive. Transitive causal chains are ones of the form $X \rightarrow Y \rightarrow Z$, and in which $X$ increases the probability of $Y$, $Y$ increases the probability of $Z$, but $X$ paradoxically *decreases* the

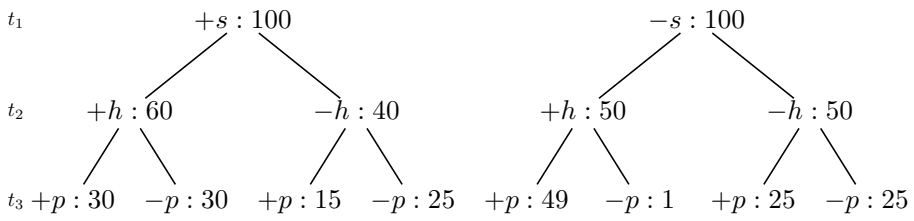


Figure 6.5: A partitioning of the individuals who smoke ($+s$) or don't smoke ($-s$), had a heart attack ($+h$) or didn't ($-h$), and who suffered chest pains ($+p$) or didn't ($-p$), with the observed size of each partition. Adapted from Eells & Sober (1983).

probability of $Z$[9].

## Intransitive Degenerate Chains

Eells & Sober (1983) have demonstrated that the Markov condition implies transitivity, and so intransitive causal chains violate the Markov condition. The reason is that, for example, where $Y$ increases the probability of $Z$, $X$ decreases the probability of $Z$, therefore knowing $X$ yields additional information about $Z$ beyond knowing $Y$ alone. Eells & Sober give the following example of an intransitive case.

Suppose that we observe 200 individuals, and observe their smoking behavior $S$ at time $t_1$, the occurrence of heart attacks $H$ between times $t_1$ and $t_2$, and the occurrence of chest pain $P$ between times $t_2$ and $t_3$. Individuals may smoke ($S = +s$) or not ($S = -s$), have a heart attack ($H = +h$) or not ($H = -h$), and suffer chest pains ($P = +p$) or not ($P = -p$). One possible set of observations are plotted in Figure 6.5.

Ostensibly, smoking is a cause of heart attacks, and heart attacks are a cause of chest pains: $S \rightarrow H \rightarrow P$. We can estimate the conditional probabilities from the

[9]Or, more generally, where the correlation between the first and last elements in the chain is opposite the correlation between the individual links.

observed frequencies.

$$Pr(+h|+s) = \frac{60}{100} = 0.60$$
$$Pr(+h|-s) = \frac{50}{100} = 0.50$$

Thus, smoking increases the probability of a heart attack. Moreover,

$$Pr(+p|+h) = \frac{30+49}{60+50} = 0.72$$
$$Pr(+p|-h) = \frac{15+25}{40+50} = 0.44$$

Thus, heart attacks increase the probability of chest pain. But notice that,

$$Pr(+p|+s) = \frac{30+15}{100} = 0.45$$
$$Pr(+p|-s) = \frac{25+25}{100} = 0.50$$

Smoking reduces the probability of chest pain. The causal chain is intransitive. But notice what makes the chain intransitive: The probability of suffering chest pains depends not on the occurrence of a heart attack alone, but also on whether the individual smoked: Smokers are equally likely to suffer chest pain as not following a heart attack, where non-smokers are more likely to suffer chest pain than not after a heart attack. But this means that heart attacks do not screen smoking from chest pain:

$$Pr(+p|+h) = 0.80 \neq Pr(+p|+h\&+s) = 0.58$$
$$Pr(+p|+h) = 0.80 \neq Pr(+p|+h\&-s) = 0.50$$
$$Pr(+p|-h) = 0.44 \neq Pr(+p|-h\&+s) = 0.38$$
$$Pr(+p|-h) = 0.44 \neq Pr(+p|-h\&-s) = 0.50$$

Thus, $P$ and $H$ are not conditionally independent given $S$, and hence $H$ fails to screen $P$ from $H$:

$$Pr(P|H) \neq Pr(P|H\&S)$$

Yet, **(CM)** requires that

$$Pr(P|H) = Pr(P|H\&S)$$

Thus, intransitive causal chains are degenerate.[10]

However, Intransitive Degenerate Chains are not limited to the realm of fiction. Consider the case, discussed in Sober & Lewontin (1982) of heterozygote superiority. In diploid organisms, genes come in pairs; much of the time, genes will have two variants, a dominant and a recessive. But not all pairs of genes exhibit strict dominance: Sometimes their effects are additive in some way. If one variant $G$ encodes for the production of a protein that is necessary in small quantities for survival, but that is lethal in large quantities, and the second variant $g$ does not encode for the protein, then the two homozygous pairs $GG$ and $gg$ will have lower overall fitness than the heterozygous pair $Gg$—in the first case, too much protein is produced, in the second not enough, but the heterozygous pair produces just the right amount. Cases of heterozygote superiority, then, are cases where we cannot say that gene $G$ has a high fitness, because it is only in the context of being paired with gene $g$ that fitness is high (much as how, in the made-up example, it is only heart attacks in the context of not smoking that cause a dramatic rise in chest-pain cases). Thus, although $G$ encodes for the production of a protein, and that protein is necessary for survival, transitivity fails, because it is not the case that the presence of $G$ in a population increases the average fitness. Such cases of heterozygote superiority present the possibility of Intransitive Degenerate Chains—and insofar as such mechanisms violate modularity too, we should worry.

[10]Sober does not consider the case that there might be a missing causal link in this model, that if added would render the model Markovian, because his concern is in showing that intransitivity implies a violation of the Markov condition. Positing a latent common cause to explain the degeneracy would also (presumably) eliminate the intransitivity, and hence would add nothing to his argument.

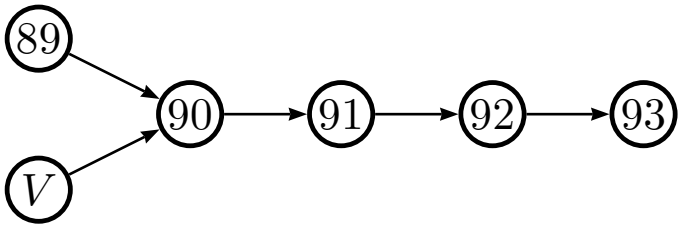


Figure 6.6: A causal model of herring populations in Prince William Sound from 1989–1993. $V$ represents the Exxon *Valdez* oil spill, and 89–93 represent the state of the herring population at each year 1989–1993, culminating in the population collapse in 1993.

## Transitive Degenerate Chains

But although intransitivity is sufficient for degeneracy (Eells & Sober, 1983), it is not necessary (Suppes, 1986). Suppes argues that some transitive causal chains violate the screening-off condition, and then offers several examples of such chains: The asymptotic behavior of certain parameterizations of the iterated prisoner's dilemma, weather forecasting, and econometric models relating consumption and disposable income. These examples share a crucial feature, in that all model causation as a relation among events (the prisoner's response at iteration $n$; the temperature this morning; disposable income this quarter), yet in these models each event depends not only upon the most proximal prior event, but upon *all* prior events. But the causal Markov condition says that an event's direct cause—namely, the most proximal event—is sufficient to predict the event with maximal precision. Knowing the full history will not add anything to our knowledge or the correctness of our prediction of that event. Thus, these examples are Degenerate Chains. A concrete example of a transitive Degenerate Chain will make the oddity of these cases clear.

As a result of the 1989 Exxon *Valdez* oil spill, researchers and fishers grew concerned about the stability of the herring population in Prince William Sound. By 1992, herring population had not shown a significant decrease in size, leading some

researchers to conclude that the oil spill had not threatened the population's stability. Yet in 1993, the population collapsed dramatically (and has yet to recover). The exact causes of the collapse—and in particular, the role of the *Valdez* spill—are still a subject of intense debate. The reason for debate is the Markovian intuition that the events of 1989 could not be a direct cause of the events of 1992 (because causation does not jump temporal gaps), yet there is no other compelling explanation for the population collapse to be found. Thorne & Thomas (2008) have presented compelling evidence that, despite the a four-year gap between the spill and the population collapse, the spill is nevertheless the best explanation of the population collapse, intervening events notwithstanding.

The problem is this. For any given year's population, its only cause is the prior year's population (see Figure 6.6. The Markov condition tells us that, to predict a given year's population, knowing the population at prior year is sufficient. However, there is nothing in the data for the herring population in 1992 that would have predicted the collapse in 1993—hence the surprise. Yet, there is nevertheless something in the data for a much earlier year, specifically the Valdez oil spill in 1989, that does not appear in the data for 1992 and that does provide additional information that could predict the collapse in 1993. Thus, the population in 1992 does not screen the population in 1991 from the oil spill in 1989, and we have a Degenerate Chain and a violation of **(CM)**.

Nevertheless, unlike the cases described above, this kind of Degenerate Chain is transitive. The reason that changes in biological populations are Degenerate is because they don't depend *only* on the previous year's population size; rather, the state of any population is influenced in addition by multi-year trends. Oversimplifying a bit to make my point, a large population last year doesn't necessarily entail a large population this year if there has been a long-term downward trend in size. In the case

of the herring in Prince William Sound, Thorne & Thomas claim that the *Valdez* spill set in motion a subtle trend over the four years following that ultimately caused the 1992 collapse, but that the trend was not apparent by looking at any single year's data. Thus, in the case of the Prince William Sound herring, the causal chain from the Valdez spill to the population collapse was transitive: The spill is clearly an indirect cause of the collapse.

What hay are we to make of Degenerate Chains? Degenerate Chains violate the screening-off condition in a different way than do Polluting Factories. But where Polluting Factories do not necessarily run afoul of **(PM)**, I will now argue that Degenerate Chains do necessarily violate **(PM)**.

## 6.4 Degenerate Chains and (PM)

Degenerate Chains violate **(CM)**. And, much like Polluting Factories, they can violate **(PM)**, in virtue of violating **(PMb)**. The way in which Degenerate Chains violate or satisfy **(PMb)** is particularly illuminating: A Degenerate Chain $A \rightarrow B \rightarrow C$ that *satisfies* **(PMb)** does so when the degeneracy disappears during interventions into $A$. Yet, part of the intuition behind modularity is that interventions should not alter the causal structure (or at least circumscribed parts of the causal structure) of the mechanism being investigated. This tension, I will argue, is reason to abandon **(PMb)** as part of a formal articulation of modularity.

Let us first take the case where we have a Degenerate Chain $A \rightarrow B \rightarrow C$ in which the degeneracy holds during interventions into $A$. Because the chain is degenerate:

$$Pr(C|B) \neq Pr(C|B\&A)$$

And because, *ex hypothesi*, the degeneracy holds under interventions into $A$, then $B$

will fail to screen interventions into $A$ from $C$:

$$Pr(C|B) \neq Pr(C|B\&\,\mathrm{set}(A))$$

Again, recall that we can analyze **(PM)** into the conjunction of two claims, **(PMa)** and **(PMb)**:

**(PM)** Suppose a set of variables $\mathbf{V}$, and $Z, Y \in \mathbf{V}$, and $Z$ is distinct from $Y$. Then $\forall Y \forall Z$

**(PMa)** If $Y$ is a non-descendant of $Z$, then $Y \perp \mathrm{set}(Z)$.

**(PMb)** If $Y$ is a descendant of $Z$, $Y \perp \mathrm{set}(Z)|\,\mathrm{parents}(Y)$.

Because $C$ *is* a descendant of $\mathrm{set}(A)$, **(PMa)** is trivially satisfied. And, because $\mathrm{parents}(C) = \{B\}$, **(PMb)** entails that

$$Pr(C|B) = Pr(C|B\&\,\mathrm{set}(A))$$

which is false because (again, *ex hypothesi*) the chain is degenerate under interventions into $A$. So **(PMb)** is violated, and hence **(PM)** is as well.

Now, let us consider the case in which the degeneracy disappears under interventions into $A$. In this case,

$$Pr(C|B) \neq Pr(C|B\&A)$$

but, *ex hypothesi*, the degeneracy disappears under interventions into $A$,

$$Pr(C|B) = Pr(C|B\&\,\mathrm{set}(A))$$

Which is not a violation of **(PMb)**, and so **(PM)** is satisfied.

These two results are quite odd, as I will now argue. One might read the intuitive notion of modularity as I developed it in Chapter 5 as the idea that interventions should not disturb the causal structure downstream of the intervention. But,

Hausman and Woodward's articulation **(PM)** embodies a more nuanced idea: That interventions should not disturb the probability distribution downstream of the intervention. Where **(PMa)** is a principle that allows us to determine which other components are effects of an intervened-upon variable, **(PMb)** is a principle that allows us to sort those effects as *direct* or *indirect* by inspecting the probability distribution over those effects. Thus, in this guise, **(PMb)** can be read as a principle that prohibits interventions from altering the probability distribution downstream of the intervention—presuming that the causal system already satisfies the causal Markov condition.

This rider is not idle; it brings out what is odd about **(PMb)** in the context of Degenerate Chains. Non-Markovian systems—Polluting Factories and Degenerate Chains—satisfy **(PMb)** precisely when interventions *do* alter the probability distribution downstream of the intervention. Yet, **(PMb)** is supposed to ensure that interventions do *not* alter the probability distribution downstream of the intervention.

The tension is readily resolved when we see that **(PMb)** is a very specific requirement that is only expected to hold of Markovian systems. **(PM)**, in other words, is a modularity principle only for Markovian causal systems. But if we want a maximally broad formulation of modularity, one that does not presume a causal system will satisfy the causal Markov condition, **(PMb)** cannot be a proper part of our articulation of modularity.

If the above argument is right, the existence of Degenerate Chains (and indeed Polluting Factories) stands as a good reason for abandoning **(PMb)** as part of our analysis of modularity. But at what cost? What are the limitations of causal inference from **(PMa)**, without **(PMb)**? We can only well and truly abandon **(PMb)** once we are certain that what remains is a tractable modularity principle.

## The Cost of Eliminating (PMb)

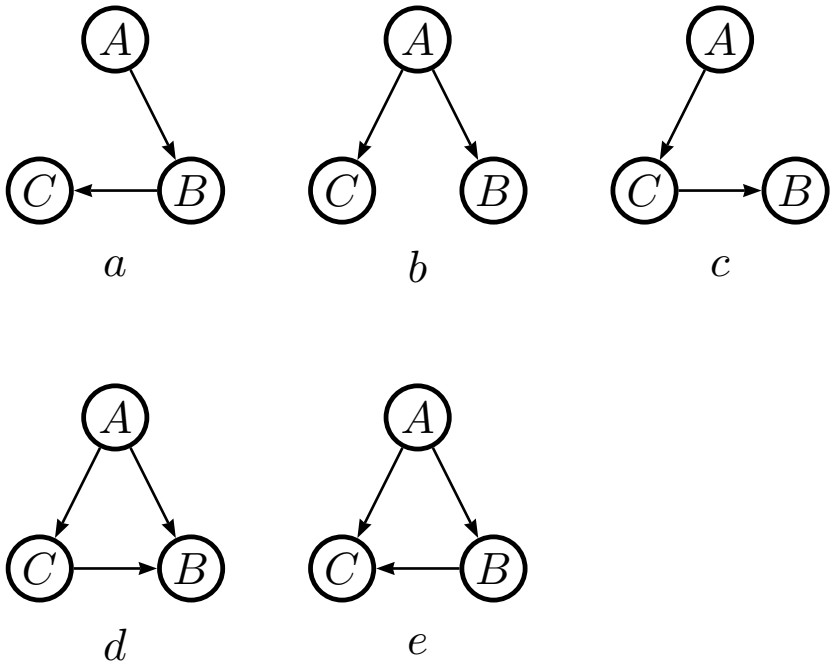


Figure 6.7: Modularity equivalence class for interventions into $A$ that result in observed correlations with $B$ and $C$.

Consider the causal structures in Figure 6.7. Suppose that the true structure is represented by 6.7.a, and we intervene into $A$. We observe that both $B$ and $C$ correlate with our intervention. If we assume only **(PMa)**, which claims interventions into $A$ will render it independent of its non-descendants, we can only conclude that $B$ and $C$ must be descendants of $A$: We cannot, on the basis of that single intervention and **(PMa)**, discern which of the structures in Figure 6.7 is the true structure.[11,12] Narrowing the equivalence class down requires either additional interventions or ad-

[11]This claim assumes cycles are not permitted; if we admit cycles, the equivalence class grows larger.

[12]Notice that the equivalence class in Figure 6.7 is *not* a Markov equivalence class, because we are not assuming that the causal structure will satisfy the Markov condition. Instead, I call it a modularity equivalence class because it contains those structures indistinguishable on the assumption that they satisfy modularity (in the form of **(PMa)**).

ditional assumptions or both. Let us intervene some more.

If we additionally intervene into $C$, we will find no correlation with $B$ (because in the true causal structure, $B$ is not a descendant of $C$). But **(PMa)** only permits us to draw conclusions on the basis of observed correlations: A finding of no correlation does not rule 6.7.c and 6.7.d out of the equivalence class. An intervention into $B$ will reveal a correlation with $C$, and will thus permit us to rule out 6.7.b, and with the additional assumption of acyclicity, 6.7.c and 6.7.d as well. Yet, 6.7.a and 6.7.e remain in the class, having exhausted the possibilities for variables to intervene into. This should be unsurprising: A Degenerate Chain is degenerate precisely because a degenerate structure as in 6.7.a will exhibit precisely the same probability distribution as a non-degenerate structure as in 6.7.e. The Degenerate Chain is thorny precisely because $A$ and $C$ are spuriously correlated in just the same way as though $A$ were a direct cause of $C$, leading **(CM)** and **(PMb)** to (falsely) identify that spurious correlation as a direct causal correlation.

Of course, if we can know that the distribution over $\{A, B, C\}$ satisfies **(CM)**, then we can eliminate 6.7.e as well, because $B$ will screen $C$ from $A$ during interventions on $A$. And since this is precisely the kind of limited Markov claim made by **(PMb)**, jettisoning **(PMb)** as I am suggesting means we cannot take this additional step without an additional reason to accept **(CM)** in the particular circumstances of our experiments.

When we can't be sure that the distribution over $\{A, B, C\}$ will satisfy **(CM)**, then without **(PMb)**, we can only get as far as ruling out every case but 6.7.1 and 6.7.e. This is the inferential ability lost when we assume only **(PMa)**.

On the other hand, if we *know* that the distribution over $\{A, B, C\}$ does not satisfy **(CM)**, then we've *gained* (in some perhaps attenuated sense) inferential ability, because **(PM)** and **(CM)** (taken as assumptions individually or together) will

yield false results, where **(PMb)** yields a set of results containing the true result. If $\{A, B, C\}$ is a Degenerate Chain, and with only an intervention on $A$, **(PM)** will falsely pick out 6.7.b as the causal structure, where the set picked out by **(PMa)** will at least *contain* the true causal structure. In this case, **(PMa)** is correct, where **(PM)** is not. Consider what happens if we additionally perform an intervention on $B$. **(PM)** and **(PMa)** will both inform us that $B$ is a cause of $C$; yet, this result is inconsistent with **(PM)**'s earlier justification of the claim that $B$ is not a cause of $C$. **(PMa)** is a more cautious principle that in these circumstances at least does not result in a contradiction when we intervene on different variables at different times. (Hence the aforementioned gain in inferential ability).

So, I conclude from this discussion that the best analysis of modularity is **(PMa)**, and not **(PM)** for the reasons iterated above. First, **(PMb)** runs counter to the idea of modularity, by requiring that interventions impose changes on the probability distribution of a causal system when such does not satisfy **(CM)**. Second, when applied to causal systems that do not satisfy **(CM)**, **(PM)** yields incorrect results, where **(PMa)** does. Although we do lose some inferential ability by rejecting **(PMb)**, we can regain it by additionally assuming **(CM)** when such assumption is independently warranted.

## Why Keep (PMa)?

There are many ways that the Markov condition can be violated. Why should we hold Polluting Factories and Degenerate Chains as special cases that decide how we should craft a modularity principle? Consider an example I will call the Cosmic Prank. Suppose we have two variables $A$ and $B$, which share no causal connection between them, no shared common cause, no shared joint effects. $A$ and $B$ are independent, $Pr(B) = Pr(B|A)$. Suppose that we intervene into $A$, and find observe that $Pr(B) \neq$

$Pr(B|\,\text{set}(A))$. More prosaically, we observe that $A$ and $B$ spontaneously correlate when we intervene into $A$—as though the universe were playing a perverse joke on us. **(PMa)** is violated, and, as we are holding tenaciously to this principle, we would be led to falsely conclude that $A$ must be a cause of $B$. Statistically speaking, somewhere, such Cosmic Pranks are quite likely to occur. Why do Cosmic Pranks not count as evidence against **(PMa)** in the same way that Degenerate Chains count as evidence against **(PMb)**?

The answer to this question lies in our conception of what it means to be an intervention, discussed in the previous chapter. An intervention is a highly constrained alteration of a causal system designed to sort causal correlations from accidental correlations, in order to inform us which parts of the system are effects of the intervention. To achieve this end, interventions must disrupt certain correlations and preserve others. Which must be disrupted, and which must be preserved? Suppose we intervene into $A$. Then, as I argued in the previous chapter, any accidental correlation involving $A$ must be broken; moreover, any causal relation with $A$ as an effect must also be broken. These correlations must be broken on pain of making a false positive causal inference about $A$. Modularity is supposed to capture this intuition. Moreover, if possible, any causal relation with $A$ as a cause (directly or indirectly) must be preserved, on pain of making a false negative causal inference about $A$. Invariance is suppose to capture this intuition, although Woodward's expressions of modularity are meant to encapsulate invariance claims as well.

Degenerate Chains present an interesting test case to this last requirement; they present a case where, in a chain $A \to B \to C$, $C$ depends on $A$ in a way that violates some common-sense notions about causal correlations, but that is nevertheless causal in character. Expecting an intervention into $A$ to impose the Markov condition on its descendants seems to ask that an intervention disrupt certain causal correlations

in which $A$ is the cause. As I have shown, in the case of Degenerate Chains, this too-stringent requirement leads us to false conclusions. At any rate, the peculiar probabilistic structure of Degenerate Chains is not at odds with the intuitions about modularity expressed above.

But Cosmic Pranks fly in the face of these intuitions about what interventions are, and what they are supposed to achieve—which is why I call them pranks. The structure of the Cosmic Prank is specifically designed to contradict the intuitions above, namely that interventions are useful tools for discovering the effects of the intervened-into entity, where Degenerate Chains do not. Thus, it seems to me that where **(PMb)** gets the modular intuitions wrong, **(PMa)** is a perfectly adequate expression of these intuitions. Insofar as we are committed to the idea that interventions are a generally useful tool for causal discovery, then too are we committed to the idea that Cosmic Pranks are not a good reason to abandon **(PMa)**.

## 6.5 Conclusion

In this chapter I have argued for two broad conclusions. First, I have defended my account of the Manipulated Mechanism against the worry that non-Markovian mechanisms—which may be legion in biology—fall outside the scope of my account. Second, I have taken the first steps at defending a new articulation of modularity in the form of **(PMa)**, one that will serve the Manipulated Mechanism's broader goals of forming the foundation of an account of mechanistic explanation and mechanism discovery.

I have argued first that Hausman & Woodward's (1999) formulation of modularity **(PM)** does not entail **(CM)**, through the use of two counter-examples, Cartwright's Polluting Factories, and my own Degenerate Chains. Both of these kinds of counter-

example are specifically designed to violate some aspect of **(CM)**, yet there are circumstances in which either may yet satisfy **(PM)**. In both counter-example, when they violate **(PM)**, it is in virtue of violating one of its components, **(PMb)**. It would seem, then, that **(PMb)** is bearing quite a lot of weight in Hausman and Woodward's argument, as it appears to make a limited kind of Markovian claim.

But, if we are aiming for a broad and inclusive articulation of modularity to serve as the foundation for evaluating mechanistic explanations and for guiding the discovery of mechanisms in biology, then **(PMb)** looks overly specialized for our purposes. If we want to be able to account for non-Markovian mechanisms, then we must not include Markov-like claims in our account of modularity. Instead, I believe that **(PMa)** is a far better candidate articulation.

But all is not well. A new line of attack claims that even the weakened articulation **(PMa)** is sometimes violated by different class of mechanisms. In the next chapter, I consider the claim that modularity requires that mechanism components exhibit a feature that I call *modular independence*—that it be possible to surgically remove or alter a component without disrupting the remainder of the mechanism. Yet, biology is full of mechanisms that do not exhibit modular independence, and hence even **(PMa)** is regularly violated. Again, such an argument, if successful, would be quite damaging to my account of the Manipulated Mechanism and so demands our attention.

# Chapter 7

# Modularity and Modular Independence

One concern about modularity is that it seems to require that we be able to get a hold of each component in a mechanism independently of the others, that we be able to surgically intervene on that component. Call this requirement *independent manipulability*. However, the requirement of independent manipulability is quite strong (Cartwright, 2001, 2004; Chemero & Silberstein, 2008), and is rarely satisfied because many mechanisms comprise components that are tightly connected, or interact in non-additive ways. For example, components that work together in a feedback loop, such as in the action potential in the neuron, do not appear to exhibit this independent manipulability, because there is no way to intervene into one without also disturbing the other (*via* the feedback loop). In this chapter, I examine three different conceptions of modular independence, each with a corresponding argument that that form of modular independence is rare, and hence that real mechanisms rarely satisfy modularity. I defend modularity against these claims by showing that modular independence is a too-narrow view of modularity. Each of these arguments

presupposes that modularity must be some actual feature of a mechanism. I argue that modularity is in fact a modal notion that constrains how a mechanism should behave were it being intervened into, and thus that modularity is compatible with the idea that we can use interventions to modify a mechanism to produce the required independencies. Thus, modular independence, I conclude, is not a necessary condition for modularity: When surgical interventions are not possible, we can nevertheless *create* the necessary conditions for modularity, and carry out an experiment *as if* we were making a surgical intervention. In the following chapter, I explore this idea of as-if modular independence in more detail.

One criticism of modularity is that it requires mechanisms exhibit a feature I call *modular independence*—that each of the components in the mechanism has some way of intervening into it without disturbing the remainder of the mechanism. This criticism observes that this is a very strong condition, and that very few mechanisms will have this feature. Therefore, precious few mechanisms are modular. Consider the following argument:

1. Modularity requires that the components in a mechanism exhibit modular independence.

2. Very few mechanisms (especially in biology) comprise modularly independent components.

3. Therefore, modularity is violated by these mechanisms.

4. The Manipulated Mechanism requires modularity.

5. Therefore, the Manipulated Mechanism cannot account for these kind of mechanisms

6. Scientists are nevertheless capable of reasoning about such mechanisms.

7. Nor can the Manipulated Mechanism account for the kind of inference scientists use for reasoning about them.

Premises one and two have been argued for, in various guises, by Cartwright (2001, 2004); Chemero & Silberstein (2008). Line three follows by *modus tollens*. Line five follows from three and four by *modus ponens*. I want to resist the conclusion in line five (and in the next chapter, I will return to this argument to resist the conclusion in line seven); I will do so by arguing that modularity does not in fact require that the components in a mechanism exhibit modular independence.

Critics of modularity claim that modularity requires that we be able to get ahold of each component in a mechanism independently of the others. Getting an independent hold on a component depends on whether the component has one or more properties that we can exploit for these purposes, a kind of property I call "modular independence". But, the ability to intervene into a component independently is a very strong condition, because real (as opposed to made-up or toy) mechanisms often comprise components that are tightly connected, or interact in non-additive ways. For example, components that work together in a feedback loop, such as in the action potential in the neuron, do not appear to exhibit this independent manipulability. And even when the components in a mechanism can be gotten ahold of independently, we might not know how, or doing so might be unethical (*e.g.* vivisection). Without modular independence, the argument goes, we cannot secure modularity, and thus cannot draw justified inferences using modularity. So much the worse for my account of the Manipulated Mechanism.

I begin the chapter by examining two different forms of modular independence, affordances and independently-disruptable mechanisms, and one property that can defeat modular independence, dynamiticty.

An affordance, according to Cartwright (2001), is a kind of special "handle" by which we can take the component through its full range of values independently of whatever is causing it, by overriding its causes. I call these things affordances because they have a certain affinity with Gibsonian affordances (Gibson, 1977), although strictly speaking the two concepts are different. But I do mean that they are a kind of property that offers itself up as a way of directly manipulating the component. Unsurprisingly, Cartwright asks us to note that such affordances are quite rare. I present her argument from affordances in §7.1.

Independently disruptable processes, according to Cartwright (2004) are causal links between mechanism components[1] that can be severed or altered independently of each other. Cartwright contrasts these with tightly-coupled systems, which are systems whose causal relations are overlapping or closely bound up such that the components cannot be excised from their mechanism without disrupting the mechanism more generally. Independently disruptable processes are likewise quite rare. I present her argument from independently disruptable processes in §7.2

Chemero & Silberstein (2008) provide a different perspective on independent manipulability, by arguing that dynamical systems cannot comprise components that exhibit any kind of independent manipulability, in part because there is no clear way to decompose dynamical systems into distinct components in the first place. Thus, by their lights, only linear, additive systems (*i.e.* non-dynamical systems) can exhibit independent manipulability. I present their argument in §7.3.

[1]Recall that mechanisms are hierarchical; the causal connections within a mechanism are themselves mechanisms, albeit at a lower mechanistic level. To avoid any confusion, I refer to the submechanisms within a given mechanism as 'processes'.

But, I will argue, these views of modularity are too narrow. In §7.4 I will argue that each of these argument rests on the assumption that the independence that modularity requires from an intervention must have existed within the mechanism prior to the intervention. This is a common-sensical view: How can I surgically excise this component if it is so closely bound up with its neighbors? All of the arguments make the case that the independence required at the time of intervention is somewhere latent within the intervened-into component, and that modularity is satisfied when the mechanism manifests this latent feature. And each argument is right to point out that expecting this to be a general feature of the world is overly optimistic. This common-sensical understanding is quite mistaken, however. A close reading of **(PMa)**, the formalization of modularity I advocated for in the previous chapter, reveals no such implication—there is nothing in **(PMa)** that requires an intervention to work by manifesting a latent independence, because **(PMa)** is silent about the probability distributions exhibited by a mechanism when it is not being intervened into.

Indeed, **(PMa)** is a modal concept, a principle that tells us what would be the case were we to intervene. And, when properly understood as a modal claim, and not a claim about the actual world, modularity does not require that the actual world be any particular way at all. Indeed, modularity is compatible with the idea that our interventions modify the mechanism being investigted to *create* the necessary independencies; we can use interventions to create conditions that satisfy modularity.

What does it mean to *create* an independence, to create the necessary conditions for modularity? And what are the limits to how we can modify a mechanism and still draw inferences from our interventions? I turn to answer this question in the following chapter, by examining a series of exemplar case studies where researchers have done precisely this.

## 7.1 Affordance Modularity

Interventions can work on a component by either actively cutting off its causes—what (Eberhardt & Scheines, 2007) (in a later commentary on the distinction) call a 'hard' intervention—, or by influencing the component without cutting off its causes—what Eberhardt & Scheines call a 'soft' intervention. Cartwright (2001, 2004) seeks to clarify what modularity entails under each of these conceptions of an intervention, to show that in either case modularity demands too much of the world. I begin with her work with soft interventions.

When we are using soft interventions, a modular system will be one in which each component "has a cause all of its own that can contribute to whatever its other causes are doing to make the effect take any value in its range" Cartwright (2001, p. 66). The availability of these causes—which I call *affordances*—ensures that we can get at each component without worrying about whether our intervention will affect other unrelated components, so long as we intervene through the affordance only. Call this kind of modularity *affordance-modularity.*

But why should we think, she asks, that affordances will be a general feature of causal systems? Cartwright (2000, 2001) claims that although some causal systems will surely offer affordances for manipulating them, these systems are rare.

Cartwright calls causal systems *epistemically convenient* when they are linear and deterministic, and when each effect (variable) in the system "has a cause all of its own that can contribute to whatever its other causes are doing to make the effect take any value in its range" (2001, p. 66) (what I will call *affordances*). An *epistemically convenient linear deterministic system* is a system that can be modeled using a system

of linear equations of the following form:

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= a_{21}X_1 + U_2 \\ &\vdots \\ X_n &= \Sigma a_{nj}X_j + U_n \end{aligned}$$

with a probability measure $P$ over the terms $U_1, U_2, \ldots U_n$ such that there are no cross-restraints among the various $U$-terms, each $U$-term is probabilistically independent of every other $U$-term, and no $U$-term can take the value of 0 with probability 1.[2]

The $U$-terms are what I have been calling affordances. I call them such because they provide a kind of handle by which we can get a hold of each effect in the system independently. If we are interested to know whether $X_i$ causes $X_j$, Cartwright observes, we need only hold all of the affordances steady, save $U_i$. We can then use affordance $U_i$ to vary $X_i$; and if $X_j$ varies in train, then we know not only that $X_i$ is one of $X_j$'s causes, we can also map out the specific functional relationship that holds between the two. The independence of each affordance from the rest is of particular importance for securing independent manipulability. If two affordances were correlated, then we would be unable to discern whether any changes in the dependent variable of our experiment were due to our intervention, or the correlated affordance.

This is simply the most straightforward discovery method; there are others that work for epistemically convenient systems. What makes these systems convenient is that there are easily proved theorems (which Cartwright takes the time to prove for

[2]I do not think there is anything significant about the fact that $X_1$ is the only exogenous non-affordance, and that each $X_n$ is a function over all and only $X_{n-1}, X_{n-2}, \ldots, X_1$. I am unsure why Cartwright insists that linear causal models will have precisely this structure, (except, as Eberhardt has pointed out to me, that she thinks some of the coefficients $a_{ij}$ might be 0, and uses this presentation to make that point. At any rate, I should hasten to point out that nothing hangs on it.

our sake, pp. 68–ff.) which show that it is possible to prove that a particular model is a correct representation of the underlying causal structure.

But, Cartwright asks, why should we think that *every* causal system will be epistemically convenient—will have affordances for each effect? She claims that authors as Woodward don't offer modularity as a tool in an axiomatic system, but as a general description of a feature common to all causal systems. Cartwright's argument against what she calls the doctrine of universal epistemic convenience is quite simple. Cartwright offers the toaster as a counter-example to the doctrine. Within the toaster we can find a common-cause structure for which the joint effects have no affordances. And if the effects have no affordances, then we cannot get at them independently of each other, and hence modularity is violated.

Inside a toaster is a lever arm that springs into motion when the toast is done (as measured by a temperature sensor). This lever arm has two jobs: It must cut off the flow of electric current to the heating coils, and it must push up the toast-rack so as to pop the toast out of the toast-slots at the top of the toaster. Within the toaster, there is no other method of cutting off the heating coils, nor is there any other method of ejecting toast. The lever arm is the only cause of these two effects. Being the only cause of the effects, we cannot independently manipulate the toast-ejecting behavior without also necessarily affecting the heating-coil cut-off behavior, because the only way to manipulate the toast-ejecting behavior is *via* the lever arm. So the toaster is not modular. QED.

One natural response to Cartwright's argument is to ask: Well, why not simply unbolt the lever arm from the toast-rack? Can't we intervene into the toast-rack in this way? Cartwright has two responses ready. The first response is that the act of unbolting is a *hard* intervention—but affordance modularity establishes the truth-conditions for modularity when we are using soft interventions. Unbolting the lever

arm cuts (nearly literally) the toast-rack from its direct causes; a *soft* intervention, on the other hand, because it does not cut causal arrows, is not the sort of thing that will modify the toaster mechanism. The strategy of unbolting thus misses the point of the toaster counter-example: That there is no intervention that does not modify the toaster so as to render the toast-rack and the lever arm independent.

Cartwright has a second response for those that might insist that the act of unbolting is an act of helping ourselves to an affordance in the toaster, and that therefore unbolting does count as a soft intervention. In this case, the unbolting strategy requires a claim that the bolting of the two parts is itself an additional cause of the movement of the rack, one that can be treated as an affordance. Cartwright rightly thinks that this an odd thing to demand, for "to do so is to mix up causes that produce effects within the...toaster with the facts responsible for the toaster operating in the way it does; that is, to confuse the causal laws at work with the reason those are the causal laws at work." (p. 72). All this is just to say that trying to treat the bolting of the toast-rack to the lever arm as an affordance is a category mistake: Unbolting just is a hard intervention, and claims otherwise are nonsensical.

None of this is to say that we are not permitted to unbolt the toast-rack. Far from it. But once we commit ourselves to a hard intervention, a different notion of modularity is in play (one which I examine the following section). Cartwright's response is simply that we cannot use the possibility of a hard intervention as an argument against the truth-conditions for modularity under soft interventions without begging the question.

More to the point, Cartwright claims that the constraints that epistemic convenience places on a causal system seem arbitrary; the demand for epistemic convenience amounts to a claim "that it is impossible to build a bomb that cannot be defused."

> Nor can we make a deterministic device of this sort: the correct function

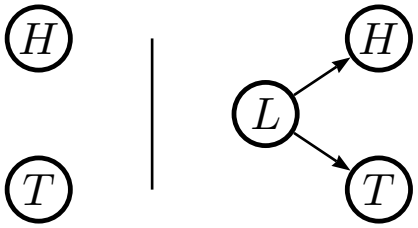


Figure 7.1: Which variables are exogenous depends on the variable set selected.

> of the mechanisms requires that they operate in a vacuum; so we seal the whole device in a vacuum in such a way that we cannot penetrate the cover to affect one cause in the chain without affecting all of them. Maybe we cannot build a device of this sort—but why not? It does not seem like the claim that we cannot build a perpetual motion machine. On the doctrine of universal epistemic convenience we either have to say that these devices are indeed impossible, or that what is going on from one step to the next inside the cover is not causation... (p. 73)

So, Cartwright concludes, the doctrine of universal epistemic convenience is not merely violated, but it is violated with regularity: the toaster is hardly an unusual system. And so modularity must not be a universal feature of the world; indeed, it will be vanishingly rare. And I certainly concede that this is a problem, if we think that the toaster represents a perfectly reasonable, indeed archetypal, mechanism, that the manipulated mechanism should be able to account for.

## Soft Interventions and Exogenous Variables—A Worry for Cartwright

I raise three issues for Cartwright's worries about modularity and epistemic convenience.

I would like to pause at this point to raise a worry peculiar to this argument, that does not bear on the remaining two. It seems that Cartwright has in mind an odd notion of intervention, that we can only intervene into exogenous causes. Why is it that we can *only* intervene via affordances? Why can we not intervene directly into the various variables directly? Whether a cause is exogenous is determined by our choice of variables.

Consider the two models of the toaster in Figure 7.1. In the left-hand model, both the movement of the toast-rack and the shutting-off of the heating coil are exogenous—they have no causes (in the set of variables used in the model). But in the right-hand model, they are not: They share a common cause, the lever-arm, which is exogenous. Neither model is more or less correct than the other, *given the choice of variables.*

But, if Cartwright's implicit claim that we can only perform soft interventions via exogenous variables is right, then the *possibility* of intervention will depend on our choice of variables. On the left-hand model, we can intervene into the toast-rack; on the right-hand model we cannot. But this cannot be right. How could our variable choice—a pragmatic, abstract choice—affect the metaphysical, causal structure of a mechanism? There is no *real* difference between the toasters in Figure 7.1, only a difference in representation. Yet, Cartwright's notion of modularity has committed her to the very puzzling result that the left-hand diagram represents a modular mechanism, and the right-hand one does not. This very puzzling feature of her notion of intervention strongly suggests that we should not put too much stock in her worries about the possibility of interventions in the case of the toaster.

## 7.2 Independently Disruptable Processes

When we are limited to soft interventions, she argued that modularity requires unique affordances for each mechanism component. Cartwright (2004) argues that when we can make use of hard interventions, modularity requires, not unique affordances, but that each causal relation consists of its own process or mechanism that is independent of the process of mechanism underlying every other causal relation in the system.

Consider again Cartwright's archetypal linear deterministic system:

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= a_{21}X_1 + U_2 \\ &\vdots \\ X_n &= \Sigma a_{nj}X_j + U_n \end{aligned}$$

Cartwright would say, when hard interventions are available, that such a system is modular, not in virtue of anything about the $U$-terms, but when for each $X_i$ it is possible to rewrite the equation as $X_i = x$ without altering any of the causal relations represented by the other equations in the system. This is clearly the notion of modularity that Woodward (2003) has in mind. Woodward calls such single replacements *wiping-out* the causal generalization for $X_i$, and setting $X_i$ to a particular value $x$.

When will it be possible to wipe out each equation independently of others? Hausman & Woodward (1999) claim that systems whose effects are the result of distinct mechanisms[3] will satisfy modularity in this way. They claim, plausibly, that

[3]Hausman & Woodward (and Cartwright) use the term 'mechanism', and I am fairly certain they mean 'mechanism' in much the same way as Glennan (1996) or (2002); but I find that their use of the term is very loose and unconstrained. I will continue their use of the term, but only with a little reluctance. Woodward, for example, often uses 'mechanism' ambiguously: to refer to that which links cause and and effect, and to refer to the equations in a causal model, which of course pick out multiple cause-effect relations, each with their own mechanism linking them together. I not make too much hay of this unfortunate equivocation. Cartwright, as we shall see, makes no such promise.

> The central presupposition... is that if two mechanisms are genuinely distinct it ought to be possible (in principle) to interfere with one without changing the other. Conversely, if there is no way, even in principle, to decouple mechanisms—to interfere with one while leaving another alone—then the mechanisms are not distinct.... This understanding of distinctness of mechanisms plus the assumption that each equation expresses a distinct mechanism implies modularity: it is, in principle, possible to intervene and to disrupt the relations expressed by each equation independently. Hausman & Woodward (1999, p. 549)).

So, on Hausman & Woodward's view, modularity requires something I will call *independently disruptable processes*[4]. These are different from affordances. Affordances are unique *causes* of each variable in a causal system that can be used as a kind of handle. Independently disruptable processes, on the other hand, are unique *causal relations* linking pairs of variables that can be broken or disrupted independently of the other causal relations.

Cartwright believes that the requirement of independently disruptable processes is too strong. Nature, Cartwright notes, has no obligation to be friendly or accessible. Indeed, the case of the Polluting Factory, introduced in the previous chapter, is one potential counter-example: chemical $X$ and pollutant $Y$ are correlated because they are produced by non-distinct mechanisms (as will it be for all cases of product and by-product).

Cartwright (2004) instead focuses on a different counter-example to the requirement of independently disruptable processes, the carburetor. The carburetor is an in-

[4]In their discussions, Cartwright and Hausman and Woodward use the term 'mechanism'. But they are in fact reffering to the mechanism that explains a particular causal relation within a higher-level mechanism. To avoid confusion, I will refer to the causal relations within a mechanism as 'processes'. Nothing weights on this bit of terminological distinction.

genious piece of engineering, which manages several functions simultaneously through the operation of a single mechanism: the Venturi effect. A venturi tube is a tube with a constriction in the middle. When fluid (such as air) passes through the tube, it speeds up as it passes through the constriction. But the Bernoulli principle states that as the speed of a fluid increases, the lateral pressure it exerts decreases. In a carburetor, an emulsion tube is placed right at the constriction such that incoming air from the air filter draws gasoline through the emulsion tube at the venturi constriction. The gasoline is atomized by the narrow emulsion tube, and is evenly distributed through the air by the time the air reaches the other side of the constriction. In this way, the carburetor can ensure that the engine receives a proper mix of air and fuel. One desirable feature of the carburetor is that the amount of gasoline drawn from the emulsion tube is proportional to the velocity of the air passing through; and since the air is drawn through the tube by the vacuum created by the movement of the engine's pistons, the amount of gasoline supplied is thus proportional to the speed of the engine.

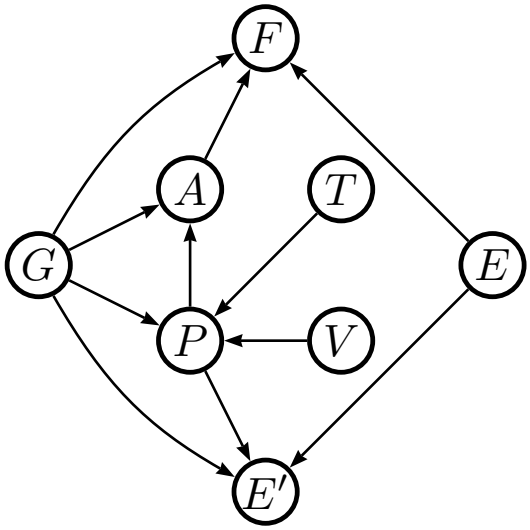


Figure 7.2: Graph of Cartwright's Carburetor

Here is Cartwright's model of the most central features of a carburetor's operation:

$$\begin{aligned} F &= f(A, E, G) \\ A &= g(P, G) \\ E' &= h(E, P, G) \\ P &= j(V, T, G) \end{aligned}$$

where $F$ is the gasoline in venturi tube; $E$ is the gasoline in the emulsion tube; $E'$ is the gasoline exiting emulsion tube; $A$ is the airflow through the venturi constriction; $P$ is the air pressure in the venturi constriction; $V$ is the vacuum pressure; $T$ is the setting of the throttle valve; and $G$ is the geometry of venturi chamber. Figure 7.2 presents a graphical version of the same model.[5]

The model succinctly describes the salient features of the operation of the carburetor described above. It also captures very neatly the central role that the particular shape of the venturi tube plays; variations in the shape of the constriction will have consequences for all aspects of the operation of the carburetor.

Notice that the only exogenous variables are $T$, the setting of the throttle value (which follows the position of the acceleration pedal in the cab); $P$, the vacuum of the engine (which is a function of the rotational velocity of the engine); and $G$, the geometry of the venturi tube. Although we can intervene into $T$ and $P$, the ingenious design of the carburetor, Cartwright notes, precludes the possibility of wiping-out any of the equations. There is no independent way to re-write the equation for, *e.g.* the airflow through the venturi without actually cracking open the carburetor and modifying the shape of the venturi:

> The conclusion is that an intervention on any of the causally determined

[5]The interpretation of the variables as parts in this model is Cartwright's own use of the Default View semantics. Since nothing I discuss here hangs on the semantics for causal models, I will not make any hay of her use of the Default View.

> variables requires an alteration to the geometry of the chamber; any such alteration will, of necessity, alter the relationships expressed in the other causal equations, in violation of modularity. Cartwright (2004, p. 23)

Notice that this is not a claim about the existence or independence of affordances. Rather, the claim is that in the carburetor there is really only one causal process at work—the Venturi effect—, and this one process is implicated in the causal relations governing several effects. I call such systems, in which one causal process links the workings of many otherwise disparate effects, *tightly coupled*. If we wanted to manipulate one effect, say the airflow, by intervening into the geometry of the venturi constriction, we are out of luck because there is no *distinct*, *independently disruptable process* linking the geometry to the airflow. We could intervene into the venturi with respect to the airflow for the purposes of testing the effects of changes in the venturi on the airflow, but only at the cost of intervening into every other variable that depends on the geometry of the venturi constriction; and so, we cannot intervene into any of these independently. Thus, the carburetor—and all tightly coupled systems, which are legion—violates modularity.

Biology presents another example: The inner ear of mammals. The inner ear mediates two important functions, hearing and balance. The cochlear system in the inner ear translates pressure waves in the air into pressure waves in a fluid that are then propagated through the cochlea, which contains many small hair-like cells (called hair cells) which move in the fluid in response to changes in pressure. The movement of these hair cells is then translated into nerve impulses. The vestibular system uses the very same fluid, but in a slightly different way. Movement of the head causes fluid to move through the vestibular labyrinth, also lined with hair cells. Again, the movement of the hair cells is translated into nerve impulses. Both the cochlear system and the vestibular system depend on the movement of the same fluids, common to

each system, and so an infection of the fluid called labyrinthitis often causes both hearing loss and dizziness as a result. Thus, not only is the vestibular fluid a common cause of hearing and balance, but the mechanism for hearing and the mechanism for balance are not distinct, in that both make use of the vestibular fluid and the inner ear. The inner ear is thus tightly coupled, in that we cannot intervene independently into either the cochlear or the vestibular system via the vestibular fluid.

One response to this kind of example is to observe that we are not limited to intervening into the airflow via the venturi geometry, or the cochlear system via the common fluid. To modify the airflow we need only mount a blower fan between the air-filter and the carburetor; then we can manipulate airflow by simply blowing more or less air into the carburetor. But such a response misses the point. This response presupposes that there really are distinct processes at play, or that we can simply arbitrarily insert new processes at will. And maybe in the case of the carburetor we can, but this is no guarantee that the strategy is universal, or even widely applicable.

Nevertheless, we might worry, with Hausman & Woodward (1999), that if there are four (correct) equations, one for each effect, there must be four distinct processes, one for each effect (and so blower-fan–type strategies are generally sound). But this worry is an illusion, Cartwright tells us, created by the structure and nature of systems of linear equations. Putting the Polluting Factory to new work, she responds:

> I have not yet figured out how to represent separate processes by separate equations. Look for instance at the probabilistic equations [for the Polluting Factory]. There is one equation for each separate effect. As I understand it, Hausman & Woodward think that the two effects [chemical $X$ and pollutant $Y$] studied in this example are not produced by distinct processes but rather by the same process. So we should have one equation, an equation for the process, rather than two. But what is this equation?

> What, for instance, are the quantities to be equated? Cartwright (2002, p. 416, footnote 6)

Although it is easy to read Cartwright's response as a certain willful refusal to be reasonable, she is raising an important point: Systems of equations are useful tools for representing the functional relationship between causes and effects, but we should be careful about how we interpret our models. In the case of a causally-interpreted system of linear equations, Hausman & Woodward reify the '=' as a causal process. But it's not clear what this reification adds to what the equations already tell us. If each term on the right-hand side is a cause of the term on the left-hand side, what more do we gain by insisting that, moreover, the equation represents a distinct process? Cartwright thinks we gain nothing, and I am inclined to agree.

What do we lose by discarding the requirement that equations represent independently disruptable processes? Hausman & Woodward's quote above is clear: We lose modularity. But why should we think that independent manipulability requires distinct process? Indeed, it is at this point that both Cartwright and her adversaries seem to lose sight of what modularity requires of the world. In the remainder of this section, I shall argue that modularity does not in fact require independently disruptable processes, that the carburetor is in fact modular, and so presents no counter-example to modularity claims.

## Modularity in Tightly-Coupled Systems

Consider the following example of a tightly coupled system. Fred and Ginger dance with each other. As each dances, their position changes (to simplify things somewhat). The dancers' position is a function of three things: the choreography of the dance, set ahead of time; error resulting from imperfections in the reproduction of

movement from the choreography; and movement in response to the partner. If Fred steps forward slightly from a slight imbalance, Ginger will step backward slightly in response.

Thus, the position of each dancer is a function of, among other things, the position of the dancer just a moment ago, and the position of the other dancer. This is somewhat ponderous, however, especially when one reflects on what 'just a moment ago' means, and the fact that there is a slight lag between the other dancer's movements and this dancer's response. It makes much more sense to talk about the difference in position of one dancer with respect to time as a function of the difference in position of the other dancer with respect to time. This, of course, is the same as talking about the velocity (difference in position) of one dancer as a function of the velocity of the other dancer.

But the very reasons that lead us to consider the positions of Fred and Ginger as partial differentials (as per the previous paragraph) are the same reasons that might lead us to reject modularity. Fred's and Ginger's movements are not distinct, because there is a single continuous reciprocal feedback mechanism connecting them—a causal process common to both Fred's movements and Ginger's. We cannot, as with the carburetor, intervene to wipe-out one generalization, say that governing Fred's movement, without disrupting the other, that governing Ginger's movement, because they are both the result of a single process linking both movements.

Cartwright's worry about tightly coupled systems is an epistemological worry, rather than the metaphysical worry she casts it as being. Just because it would be difficult if not impossible to have Ginger move in a different way without altering the relationship between Ginger and Fred does not mean that modularity does not hold of the Ginger-Fred system, because we can still understand the claim that 'Ginger's movement affects (causes) Fred's movement' means 'were Ginger to step in this-or-

that way, Fred would respond as thus.'

That modularity is a modal notion suggests a way forward. That the causal structure of Ginger and Fred does not *currently* exhibit independently disruptable processes is not reason to think that it is not modular. What if we could change out the problematic common process linking Fred's and Ginger's movements with a new pair of causal processes?

We construct a pair of anthropomorphic robots that are capable of carrying out basic dancing instructions. The robots are quite simple, in that neither needs to sense or respond to its dancing partner. We program each robot to carry out a series of maneuvers meant to test how Fred and Ginger respond to their partners, and pair Fred with one robot, and Ginger with the other. We might then instruct each robot to do a simple box-trot, but to seemingly randomly (but in reality cleverly programmed) stumble slightly this way or that. We have total control over the stumbles; we need only observe how Fred responds without worrying about trying to factor out how Ginger will counter-respond, and *vice versa*. Because we can control the stumbling of each robot independently, we can independently disrupt the processes linking Fred and Ginger with their dance partners. Thus can we, at least in principle, break a feedback loop without losing information about what makes the loop work.

Thus, although Fred and Ginger are governed by several generalizations, and although, like the carburetor, these generalizations are harnessed together (although, in this case, by a feedback loop rather than a singular central feature) and so represent what Cartwright and Hausman & Woodward would identify as a single process, we have successfully intervened. Why? We have un-harnessed the generalizations by removing the mutual dependence of Fred on Ginger and Ginger on Fred. In a very loose sense, we have replaced the singular process underwriting the entire system with a new set of independently disruptable processes. By replacing wholesale the overlap-

ping, non-distinct processes we can pull the generalizations apart, and intervene into either Fred or Ginger independently, satisfying modularity.

If we return to the carburetor, we can make a similar move as with the Fred-and-Ginger case.. If our interest is to understand the particular functional relationships within a carburetor, we ought to be able to decouple the single process governing each feature. Suppose we wish to intervene to manipulate the fuel flow from the emulsion tube. If we are to decouple the process that produces emulsified gasoline we have to decouple it from the air-flow, the pressure in the venturi tube, etc. Indeed, this is easy enough: We have already noted that we can use a blower-fan to manipulate the air-flow. For a given set of setup conditions, we measure the air-flow in an unmodified carburetor. We then modify the venturi geometry. We measure the air-flow under the same conditions, and then we use a blower-fan to eliminate any differences in the air-flow between the two carburetors. Then, we can say precisely how the shape of the venturi tube affects fuel flow, without worrying that any difference in fuel flow is due to the difference in air-flow.

Notice that in the case of the carburetor, unlike in the case of Fred and Ginger, we decouple the singular process by introducing a second process: We, in effect, intervene in multiple places. One intervention is experimental (the change in the shape of the venturi tube); the other intervention is restorative (the blower-fan). Indeed, although the case of Fred and Ginger is fanciful, this kind of reasoning is precisely how scientists proceed when faced with a tightly coupled feedback system—yet more evidence that Cartwright has illicitly narrowed the range of truth-conditions for modularity. In the next chapter, I will examine the case of Hodgkin and Huxley's experiments on the squid giant axon: Their experiments depended on an apparatus called a voltage clamp which introduces multiple simultaneous manipulations into the neural cell to break apart a tightly-coupled feedback loop in just the way described above.

Thinking of modularity as a modal concept, we can see that the concept of independently disruptable processes does not exhaust the truth-conditions for modularity. Modularity makes claims about how the world *would* (or should) *be* during an intervention, where Cartwright has interpreted it as a claim about how the world *is*. Once we begin to admit possibilities, satisfying modularity becomes quite easy, even in tightly-coupled systems, so long as we can think of ways to modify the overlapping processes. I return to this modal nature of modularity in the last section of this chapter.

But first, we must return to feedback loops, which, as I discussed above, must be teased apart before we can treat a system containing them as modular. I turn now to consider feedback in more detail, because some authors worry whether feedback loops can even be decomposed into components in the first place. If they cannot, then the idea that we can pull them apart is misguided, and this modal notion of modularity that I've introduced will be untenable.

## 7.3 Dynamical Systems and Decomposition

Dynamical systems are closely related to tightly-coupled systems. Where tightly-coupled systems comprise a number of components whose interactions overlap, dynamical systems are complex systems that cannot be readily decomposed into distinct components in the first place. Because there are no distinct components, it is hard to see how a Fred-and-Ginger like intervention can work, since that strategy depends on being able to carve up a system into individual components for testing.

Dynamical systems are of interest to us, because there are a well-understood set of mathematical tools for analyzing their complex behavior, and these have lately become popular for describing and explaining biological systems. Dynamical systems

analysis focuses on the kinematics of a system; how the system evolves over time, and how the various measurable properties change in relation to each other. One early example of dynamical systems modeling in biology is the Hodgkin and Huxley model of the action potential. This model describes the movement of potassium ($K^+$) and sodium ($Na^+$) ions across a neuron cell membrane as a function of changes in potential difference across the membrane. But the model also describes the changes in potential difference as a function of the relative concentration of ions. Changes in potential are caused by differences in ion concentrations; the movements of ions are impelled by potential differences. Hodgkin and Huxley used systems of non-linear differential equations to model this behavior (a topic I will return to below).

But, as important to biology as dynamical systems are, there is a wrinkle. Modular independence requires that a mechanism be decomposable. And decomposability, as Chemero & Silberstein (2008) argue, requires the mechanism be linear—that is, non-dynamical. A non-linear system is one whose various elements do not interact in an additive fashion to produce their effects, that is, whose elements produce interaction (in the statistical sense of the term) where the contribution of one element can be modulated by the contribution of another. For example, a positive feedback loop operates by rapidly magnifying small changes in initial conditions into exponentially larger effects, larger than the sum of the individual small changes that initiated the feedback cycle. Chemero & Silberstein (2008) think that non-linearity poses a problem for modularity because there is no way to cleanly demarcate the various causal contributions being made by each cause in the mechanism, and hence there is no way to identify any possible component as a distinct cause in the mechanism—and without distinct causes, there is nothing to be modular.

Thus, if Chemero & Silberstein are right, non-linear systems will not be in general modular, and hence when the behavior of the system is produced by a mechanism,

the Manipulated Mechanism will not be able to account for this mechanism. In this section, I will argue that modularity does not require that we be able to decompose a mechanism into independent and linear functions that can each be localized to individual components in the mechanism. I will argue moreover that modularity does not require that the various components interact in a strictly linear fashion, and that therefore non-dynamical system can be modular.

## Decomposability and Non-Linearity

Modular independence requires that a mechanism be decomposable, because modular independence is a feature of mechanism components. If we cannot decompose a mechanism into components, then there is nothing to have the property of modular independence.

Decomposition, Bechtel & Richardson (1993) tell us, is the process of taking a mechanism, analyzing it into distinct *functions*, and then localizing those functions in spatio-temporal regions of the mechanism by identifying those regions that carry out or implement each of the functions. Insofar as the decomposition is correct, then the functions will be localizable to the mechanism's components. Bechtel & Richardson worry that only linear, additive functional relations will be decomposable.

An algebraic function is linear when the terms on the left side of the '=' sign are an additive combination of the terms on the right. Additivity is simply the property of being the sum of two or more terms. Thus, a linear function is one where each of the terms are combined through addition, rather than some other operator such as multiplication or exponentiation. When two terms are non-additive, they appear as an interaction (in the statistical sense of the word), where the contribution of one term can modulate or be modulated by the contribution of a second. (For an extended discussion of linearity and additivity, please see Appendix A.)

Bechtel & Richardson tell us

> Decomposition assumes that one activity of a whole system is the product of a set of subordinate functions performed in the system. It assumes that there are but a small number of such functions... [that] are minimally interactive... [and] can be handled *additively or perhaps linearly.* (p. 23, emphasis added)

Bechtel & Richardson do not argue for this claim, but Chemero & Silberstein (2008) do. They argue that at non-linear dynamical systems will defy mechanistic explanation, because they defy decomposability. In a dynamical system,

> the more localizability and decomposition fail, the harder mechanistic explanation will be, and a high degree of non-linearity is bad news for both of these. (p. 16)

This worry parallels van Gelder's (1995), who is concerned that the mind, as a dynamical system, cannot be decomposed into modules, each with a specific functional role to play. But Chemero & Silberstein come to a stronger conclusion than does van Gelder. They conclude not merely that if the mind is a non-linear dynamical system then it won't be composed of modules, but that then it won't be composed of *anything*, because the very notion of composition doesn't make sense in a non-linear context.

Chemero & Silberstein (2008) begin their argument by allowing that a system of *linear* equations can be decomposed into subsystems. They do not explicitly define what they mean by a 'subsystem', or how to decompose a mathematical model; I have attempted here to charitably fill in these gaps.

If the goal of decomposition is to find a description of the causal relations in a mechanism, and a subdivision of the mechanism into components consistent with this

description, and we are using systems of equations to describe the causal relations, then our system of equations must meet the constraints on causal interpretations discussed in Chapter 1. Specifically, there must be one equation for each endogenous variable (a variable representing an effect), with the endogenous variable on the left-hand side and all of its causes (and none of its non-causes) on the right-hand side. Thus, I presume that Chemero & Silberstein mean that decomposing a system of equations means solving that system for each of its endogenous variables. That is, as I understand them, decomposition requires that we take a linear system of equations, and manipulate it algebraically until each endogenous variable appears as the left-hand side of one and only one equation in the system. Call each resultant solved equation a *subsystem*, and the total set of solved equations the *decomposed system*.

Here is a worked example of how I take decomposition to work. Suppose we have a linear system in $n + m$ variables, with $n$ endogenous variables $\mathbf{X} = \{x_1 \dots x_n\}$ and $m$ exogenous variables $\mathbf{Z} = \{z_1 \dots z_m\}$. Then a decomposed system will consist of $n$ solved equations of the form

$$x_i = f(\mathbf{X} \setminus x_i \ , \ \mathbf{Z})$$

where $f$ is a linear function of its operands of the form:

$$f(\mathbf{V}, \mathbf{U}) = \sum a_i v_i + \sum b_j u_j.$$

But non-linear systems, Chemero & Silberstein claim, cannot be so decomposed, because they cannot (in general) be solved in each endogenous variable. Consider a

set of variables $\mathbf{X}$, and a system of differential equations in matrix notation[6]:

$$\frac{d\mathbf{X}}{dt} = f'(\mathbf{X}) = \mathbf{F}\mathbf{X}.$$

So long as the coefficients in $\mathbf{F}$ are constants (and the coefficient for the term on the right-hand side containing the variable on the left-hand side is zero), the system is linear. But if any of the coefficients is one of the variables in $\mathbf{X}$, then the function containing that coefficient will be non-linear because it is a multiplicative combination of variables rather than a simple additive combination, and because there will be a term in one of the equations of the form $x_i x_j$, where $x_i$ is the coefficient. The resulting system will violate additivity[7], and hence the system will be non-linear.

Let us suppose that we are dealing with only non-linear systems. Such systems (as the ones discussed above) are not generally solved in all of their variables, but are only solved in the first time-derivatives of their variables. To solve a system (and hence to decompose it), we need to use the calculus to rewrite the equations in the form

$$\mathbf{X}_{t+1} = f(\mathbf{X}_t) = \mathbf{G}\mathbf{X}_t.$$

Where $\mathbf{G}$ is a new coefficient matrix for the solution, and $t$ is a subscript indicating

[6]This is a nice shorthand for the system of equations

$$\begin{aligned} \frac{dx_1}{dt} &= f_1'(\mathbf{X}) = \sum_{i=1}^{n} a_{i1}x_i \\ &\vdots \\ \frac{dx_n}{dt} &= f_n'(\mathbf{X}) = \sum_{i=1}^{n} a_{in}x_i. \end{aligned}$$

The advantages of using matrix notation should at this point be clear. Notice that in a linear system, for the equation with $x_j$ on the left-hand side, the coefficient for the term $a_{ij}x_j$ will be zero, and all other coefficients will be constants.

[7]Additivity is simply the property of being made up of additive combinations of variables, *e.g.* in $f(x, y, z) = ax + by + cz$. Non-additive systems are composed of multiplicative, exponential, or logarthimic combinations of variables, *e.g.* $g(x, y, z) = xy + y^z$, *&c.*. See Appendix A for an extended discussion of additivity and its relationship to linearity.

that $\mathbf{X}$ evolves over time; $x_t$ is the value of $x$ at time $t$, and $x_{t+1}$ the value of $x$ at the next time step.[8]

In general, however, although systems of differential equations can be solved, systems of *non-linear* differential equations cannot be solved in their variables—only very simple systems and systems that exhibit special properties have solutions. For example, systems of quadratic equations in one quadratic variable (and zero or more linear variables)—*i.e.* a system of equations of the form $0 = ax^2 + by + c$, and where in each only the $x$ term is squared—can be reduced to a single quadratic equation, which can be solved using the quadratic formula $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. But in general, there is no method for solving systems of non-linear differential equations, and most systems have no solution at all[9]. So, if decomposing a system means solving it (as I take Chemero & Silberstein to mean), then most non-linear differential systems cannot be decomposed. So far so good.

The final step is to show that these mathematical difficulties are likewise difficulties for mechanistic explanation. Chemero & Silberstein identify the decomposition of a system of solved equations with the decomposition of a causal system into independent mechanisms (which is largely how, as I have shown, Bechtel & Richardson (1993) and Hausman & Woodward (1999) characterize decomposition); and they identify the result of the mathematical decomposition (the solution of the system of equations in each variable), with the mechanisms in the causal system. They conclude that the tightly-coupled nature of dynamical systems precludes the possibility of disentangling the various entities in a mechanism that system represents; that the coupled nature of the equations mirrors the deeply entangled nature of the entities:

[8]Such a representational notation *does not* presume discrete time steps. The representation allows talk of, *e.g.* $\lim_{\epsilon \to \infty} x_{t+\epsilon}$.

[9]Just because there is no method for solving an equation does not mean there is no solution. But often there aren't any solutions to be had anyway.

> When the behaviors of the constituents of a [mathematical] system are highly coherent and correlated, the [causal] system cannot be treated even approximately as a collection of uncoupled individual parts. Instead, some particular global or nonlocal description is required, taking into account that individual constituents cannot be fully characterized without reference to larger scale structures of the system such as order parameters. (Chemero & Silberstein, 2008, p. 16)

Two equations for $x_i$ and $x_j$ are coupled if $x_i$ shows up as a coefficient in the equation of $x_j$ or *vice versa*. And in a non-linear dynamical system, such coupling will occur at least once, if not in multiple places. Chemero & Silberstein are claiming that if we cannot solve a mathematical system (because of coupling), then there is no meaningful way to talk about decomposing the physical system the mathematical system represents (for the parts must therefore also be coupled). So highly non-linear physical systems cannot be decomposed in the way that mechanistic explanation requires. And if they cannot be decomposed, then these physical systems cannot satisfy modularity. QED.

Several questions for Chemero & Silberstein come to mind. First, why should we identify the solutions to a system of equations (the mathematical decomposition) with the entities of the physical system being represented (the mechanistic decomposition)? Granted that this is a view I am imputing to Chemero & Silberstein, but it seems the only way to make sense of their claim that non-linearity defies mechanistic decomposition. Second, does non-linearity, and hence non-decomposability, really violate modularity? That is, are there other ways to satisfy modularity when this particular way fails? I turn now to consider these questions in some detail.

## Non-Linearity and Modularity

Why must we be able to solve a system of equations to identify the entities of the mechanism the system represents? At first blush, we might think that only a solved system can be treated as a causal model. Each solved variable appears on the left-hand side of one and only one equation in the solved system; giving each equation a causal interpretation yields a causal model that fully describes the causes of each variable. Although this move seems natural, a closer examination reveals that this position is unfounded: There is no *a priori* reason to think that a solved system will correctly represent the causal relations within the mechanism.

We might think that, in a particular case, the *unsolved* system may give the correct causal relations. Woodward (2003) has argued that in a linear system, at most one arrangement of the variables can represent the true causal structure among those variables, and that algebraic rearrangement does *not* preserve causal relations. Extending his argument, there is no particular reason to think that the calculus will preserve causal relations either, as the calculus is a superset of algebra, and there is no particular reason to think that a system's being non-linear provides a guarantee that mathematical manipulation will preserve the causal relations. So, if an unsolved non-linear dynamical system *does* represent the true causal structure, then the solved system *will not* represent the true causal structure. And therefore there is no reason to think that the solved system is a privileged representation of causal interactions.

Consider the following non-linear causal model:

$$\begin{aligned} \frac{dZ}{dt} &= U_Z \\ \frac{dX}{dt} &= YZ + U_X \\ \frac{dY}{dt} &= XZ + U_Y \end{aligned}$$

The model is non-linear because the first time-derivatives of $X$ and $Y$ are non-additive

combinations of each other. Moreover, since the model is dynamical, it is *not* solved in its variables—rather, it is solved in the variables' first time-derivatives. Suppose that this model represents the true causal structure among the first time-derivatives of the variables $X$, $Y$, and $Z$. That is, the model represents that the velocity of each variable (it's first time derivative) is caused by the multiplicative combination of the position of two other variables.[10] There is nothing in the *mathematical structure* of the model that precludes us from setting any of the variables, *e.g.* wiping-out the equation above for $\frac{dY}{dt}$ and re-writing it as

$$\frac{dY}{dt} = \text{set}(Y).$$

This new equation represents an intervention into the velocity of $Y$. Such an intervention breaks the arrows leading into the velocity of $Y$, which include arrows from $Z$ and $X$. Interventions work by disrupting certain of a mechanism's activities. But notice that the system's being non-linear doesn't have any bearing on whether could make such an intervention. And the system's not being solved is likewise irrelevant to the metaphysical possibility of intervention. The final equations claims that $x$ is a cause of $\frac{dY}{dt}$, and says *nothing* about whether we can intervene into $\frac{dY}{dt}$ directly; the second that $\frac{dX}{dt}$ is an effect of $Y$. So, I do not think that there is any reason to think that non-linearity precludes the possibility of independent manipulability of the variables.

Neither non-linearity nor being dynamical (nor their combination) need violate modularity. On the one hand, there is no obstacle to our being able to make sense of interventions into non-linear systems. A non-linear system is one in which one or more variables appear as coefficients in the right-hand side of one or more equa-

[10] Although being a cause of velocity might sound exotic, or perhaps even untenable, consider that the mass of our sun has a concrete causal influence on not only the position of the earth, not only its velocity, but on its second time-derivative: its acceleration. So such a causal model is hardly unusual.

tions. But Chemero & Silberstein appear to conflate the modular independence with additivity. Additivity ensures that the terms in a linear function are independent, in that there are no interactions among the variables in the *statistical* sense of the word. The existence of multiplicative effects in a system means that we cannot use linear statistical methods for analyzing such systems (*e.g.* ANOVA), without additional assumptions. But for causal modeling, these kinds of interactions are innocuous. Chemero & Silberstein (2008) appear to conflate this statistical sense of interaction with the probabilistic notion of dependence, which is quite different. Two variables $X$ and $Y$ are dependent when they are correlated; they interact when they multiply together to produce a joint effect $Z$. Dependence is orthogonal to linearity: $X$ and $Y$ can correlate, yet still jointly produce $Z$ in additive fashion. Likewise could they produce $Z$ in a non-linear fashion without themselves being correlated. So, although modularity does require certain patterns of independence among the variables in a model, it hardly requires linearity. To take a quick example, although genotype and environment might interact in a non-additive fashion to produce a phenotype (Tabery, 2009), we can yet intervene on either genotype or environment without difficulty.

On the other hand, there is also no obstacle to our ability to decompose non-linear mechanisms, because physical decomposition is not isomorphic with mathematical decomposition, and so a failure of mathematical decomposability does not entail that we cannot find a mechanistic decomposition. Chemero & Silberstein (2008) appear to presume that variables in mathematical systems must represent entities, a view that I rejected in Chapter 4. If variables need not represent entities, then we need not worry that decomposing a mechanism into its entities requires that we be able to decompose or solve the system of equations that model it. The functional structure of a causal system is captured in the system of equations that best fit those functions; thus, having a system of equations that can be given a physical interpretation—whether

they are solved or not—is sufficient for decomposition.

Cartwright's argument from affordances, her argument from independently disruptable processes, and Chemero & Silberstein's (2008) argument from decomposability share a common thread. Each argument rests on a presupposition that modularity is an actual (in the sense of non-modal) feature of a causal system. But, as I turn now to argue, modularity is a modal claim about interventions, and does not depend on a causal system having any of the properties described so far in the chapter.

## 7.4 Sufficient Conditions for (PMa)

Each of the arguments I have considered in this chapter takes as their goal to undermine modularity as a general principle for causal inference. Each proceeds by demonstrating a set of conditions that must be met by a mechanism before it will satisfy modularity (it must have affordances, or must comprise independently disruptable processes, or must contain only linear, non-dynamical interactions). Then, each argument argues that these conditions are in some way extraordinary by offering some kind of counter-example: a mechanism that is seemingly commonplace, that fails the stringent conditions described, and that is yet clearly the kind of mechanism amenable to experimental analysis. Thus, each argument concludes, modularity is the wrong way to think about causal inference in these kinds of cases.

If the examples and arguments succeed, then insofar as the examples *are* mechanisms, then the manipulated mechanism will fail to account for them. For my account to succeed, I must show that, although such stringent conditions are sufficient for modularity, they are not necessary for satisfying modularity.

What is common to each argument is this. Because modularity makes a claim that certain independencies must hold during an intervention, each of the conditions

offered presume that those independencies must be the result of some latent feature of the mechanism intervened into. Affordances are a kind of 'handle' that each component must have, that we can take advantage of during a soft intervention to assure modularity. Independently disruptable processes are causal relations that have no overlap or common component. Non-dynamiticity ensures that the components in a mechanism depend only on the value of a manipulation, and not its rate of change.

But this kind of requirement is too strong. Modularity is a modal notion: it does not explicitly demand anything of the mechanism being intervened into *except during an intervention*. Modularity encodes a set of counter-factual claims about how the mechanism should behave were we intervening into it ideally or surgically. Consider again my offered formulation, **(PMa)**:

**(PMa)** Suppose a set of variables $\mathbf{V}$, and $Z, Y \in \mathbf{V}$, and $Z$ is distinct from $Y$. Then $\forall Y \forall Z$, If $Y$ is a non-descendant of $Z$, then $Y \perp \text{set}(Z)$.

We can talk about a mechanism prior to and during an intervention; **(PMa)** is a principle for linking inferences about the former to conclusions about the latter. **(PMa)** is a conditional, with the antecedent a proposition about the mechanism prior to intervention ($Z$ is not a cause of $Y$) and the consequent a proposition about the mechanism during the intervention ($Y \perp \text{set}(Z)$). In this way, it licenses inference about the causal structure of mechanism prior to intervention from observations of dependence made during an intervention: If we observe a dependency between $Y$ and set($Z$), then we can infer that $Z$ (*prior* to the intervention!) must be a cause of $Y$. But this is all **(PMa)** has to say about the mechanism prior to intervention. In particular, although **(PMa)** does not make any claims about how the independencies that must hold during an intervention should come about. It does not prohibit, for example, modification of a mechanism that does not disrupt the causal structure or

independencies required.

If, however, the authors surveyed in this chapter are right, that satisfying modularity really does require that the components in a mechanism be already independent, then they render modularity trivial—if the variables are already independent, then the inferential work is being borne by the requirement of ideal interventions. If $Z$ and $Y$ are already independent in the right way, and we are presuming our interventions to be ideal, then modularity falls out of these conditions trivially. Thus, the arguments that I've considered are merely arguments that many if not most mechanisms will not satisfy modularity *trivially* (because the components of most mechanisms are not already independent in the right way). This is hardly news.

But why should we think that modularity demands that we leave the mechanism of interest intact? Why not modify it? Cartwright (2004), as discussed earlier, has stated that we cannot inquire about the causal structure of a toaster by modifying it, because the resulting mechanism will be quite different from the original toaster, and no similarity metric will ultimately justify inference to the causal structure of the original. But, as I've just noted, modularity *just is* a principle that licenses such analogical reasoning from a mechanism being intervened into, to conclusions about the mechanism prior to intervention. In which case, her worries about modifying the mechanism seem misplaced. More to the point, biologists frequently do modify—sometimes quite heavily—the mechanisms under study. And, indeed, such modification appears licensed by **(PMa)**, so long as it is done in the service of securing the necessary independencies, and so long as it is done without rearranging the causal relations beyond what is strictly necessary for an ideal intervention (*i.e.* that we sever all and only those causal relations of which the manipulated variable is an effect). In these cases, modularity is actually doing quite a lot of inferential work not borne by the concept of an ideal intervention: **(PMa)** gives us, first, the constraints on how

much or in what ways we are permitted to modify the mechanism, and second, having modified it how to infer a causal connection in the original mechanism.

In the next chapter, I turn to consider historical cases in which biologists and neuroscientists did in fact modify the mechanism under investigation, and I show how that these modifications were performed in the service of satisfying modularity. With a handful of empirical case studies in hand, I develop a general set of constraints that modularity places on the modification of mechanisms via intervention.

## 7.5 Conclusion

In this chapter, I have considered arguments that many mechanisms violate modularity, because the necessary conditions for modularity are too strong. Weak interventions, according to Cartwright, necessitate that each effect in a system have a unique affordance for modularity to be satisfied. Strong interventions necessitate that each effect be brought about by a distinct mechanism for modularity to be satisfied. Chemero and Silberstein argued that mechanisms must be linear and non-dynamical to satisfy modularity. I have argued in each case that the author has a too-narrow view of modularity: That, properly understood, all of these conditions are indeed sufficient for modularity, but affordances, distinct mechanisms, and linearity are hardly the *only* sufficient conditions for modularity. Each of these accounts presumes that modularity demands that we not modify the mechanism under study. Yet, there is nothing in the modularity principle that places any such requirements on our investigations. Modularity does place constraints on how and when we can modify a mechanism and still derive true causal conclusions, but these constraints are not such that we can *never* modify a mechanism.

What remains is to spell out in detail a principled account of the necessary con-

ditions for modularity. I turn now to complete this task.

# Chapter 8

# As-If Modular Independence

In this chapter I elaborate on how one can satisfy modularity without requiring that the mechanism already exhibit some kind of independence—a feature I call 'modular independence'. In the previous chapter, I considered three different ways that a mechanism can exhibit modular independence: via unique affordances, independently disruptable mechanisms, and linearity, and I argued that satisfying modularity does not require that a mechanism exhibit modular independence. How can we intervene into a mechanism that does not exhibit modular independence? In this chapter, I examine three historical cases in which experimenters successfully intervened into mechanisms that did not exhibit modular independence. In each case, the experimenters had to somehow *create* the necessary independencies for satisfying modularity. I develop, on the basis of these cases, a general account of what I call 'as-if modular independence', a set of conditions on how an intervention can modify a mechanism to create the necessary independencies to satisfy modularity.

One way to satisfy modularity is when a mechanism exhibits a feature that I call *modular independence*. A modularly independent mechanism is one that has a latent structure—affordances, additive effects, *&c.*—that can be exploited by interventions so as to bring about the independencies necessary to satisfy modularity. In the previous chapter, I argued that modular independence is not necessary for modularity. Instead, I suggested that, rather than relying on some latent structure in the mechanism, we can use interventions to *create* a kind of 'as-if' modular independence, and so secure modularity without relying on actual modular independence.

In this chapter, I extend that argument by presenting three historical cases in which researchers sought to investigate mechanisms that did not exhibit modular independence. In each case, the researchers used a complex intervention that affected multiple components simultaneously in order to create as-if modular independence in these mechanisms. I show that the curious interventions used were crafted specifically—if tacitly—to secure modularity. I will argue in each case that the use of a simple, naïve intervention would result in a violation of **(PMa)**, but, in contrast, that the complex interventions actually used secured the satisfaction of **(PMa)**.

Hodgkin and Huxley, investigating the electrical characteristics of the squid giant axon, found that the voltage and the current across the cell-membrane were not only physically impossible to tease apart (being, effectively, two measurements of the same phenomenon), but formed part of a feedback cycle which could not, therefore, be broken. In §8.1 I show how Hodgkin and Huxley were able to satisfy modularity through the combined use of multiple simultaneous manipulations and negative feedback. If they had tried to intervene by simply manipulating the membrane voltage, they would have found that this manipulation would drive the membrane current up—and hence the membrane voltage too, rendering their manipulation and their measurement of the current invalid. Instead, as I will show in §8.1 they performed two simultaneous

manipulations, of the membrane voltage *and* of the membrane current, such that the axon behaved as though the voltage and current did not feed back upon each other. In this way, they could vary the voltage independently of the current, and satisfy modularity.

Otto Loewi sought to discover whether a mysterious and possibly fictious chemical substance—*vagusstoff*—was the mechanism by which the vagus nerve signaled the heart muscles. But, having simply postulated its existence, he knew nothing of its composition or how to synthesize it, except by direct stimulation of the vagus nerve. But since stimulating the vagus nerve causes the heart rate to slow, such an intervention would be unrevealing if not question-begging. *Vagusstoff*, being a mystery, could not be intervened into independently of the vagus nerve. In §8.2, I will show how Loewi surmounted this difficulty by preparing an analog that comprised two heart preparations. In the first, he stimulated the vagus nerve, and collected the fluid from its base. In the second, he stripped off the vagus nerve, and introduced the fluid collected from the first. Thus could Loewi manipulate *vagusstoff* using only stimulation of the vagus nerve without risk of violating modularity.

Pain is a difficult concept to quantify and measure; as a result, the testing of analgesic drugs can be quite a complicated affair. First, as a subjective experience, pain can only be measured indirectly; one popular model for pain measurement is the 'cold pressor' task, first identified by Wolf & Hardy (1941).[1] In this task, subjects are asked to immerse their hand in a bath of very cold water. The cold water brings about a dull, aching pain in the subject; the length of time that a subject is willing to maintain the submersion is taken as a measure of *pain threshold*. This task, offering an operational measurement of pain, appears ideal for testing the effectiveness of analgesics. Two complications arise, however. First, pain tolerance is affected by a

[1] As an interesting philosophical aside, Wolf, in this study, was the first to observe that aching 'cold' pain signals are conducted by "the small, non-mylenated [fibers] of class C" (p.531).

great many factors, all of which contribute to the measured task time; these must be accounted for and subtracted out. Second, analgesics are double-acting: They (ideally) have a pharmacological action, but they also have a psychological action known as the placebo response, which is completely independent of the pharmacological agent, and can be elicited by any substance that the patient believes to be an analgesic. The risk to modularity is that we cannot intervene ideally on the neurological factors affecting pain threshold; we have only ham-fisted interventions available to us, and as a result, we cannot attribute the sum of a subject's pain response to the drug alone. We need a way to hold the other factors, including the placebo response, steady. In §8.3, I show how the method of subtraction combined with experimental controls permit us to treat otherwise ham-fisted interventions as surgical.

In all three of these cases, researchers sought to understand a mechanism by intervening into it—and in all three cases, the mechanism defied straightforward intervention, because the experimenters could not get an independent hold on the intervention variable. Modularity, however, seems to require that the components of a mechanism be such that we *could* get a hold of each component independently. Therefore, the mechanisms above appear to violate modularity. Yet, in all three cases, the researchers were successful in their experiments. How can this be? How can a mechanism into which we cannot intervene straightforwardly possibly be modular?

In this chapter, I will argue that in each of the above cases, the researchers deployed complex or indirect interventional techniques that permitted the satisfaction of modularity. When the usual methods of carving—perhaps literally—a component from its context cannot be used (and hence, where modularity cannot be satisfied trivially), scientists often turn to techniques such as multiple simultaneous manipulations, experimentation on analogs or models, and the method of subtraction with experimental controls.

## 8.1 Multiple Manipulations

Let me begin by drawing a useful distinction. An ideal *intervention* simultaneously fixes the value of one variable[2] in such a way as to render that variable independent of its non-descendants. A *manipulation*, on the other hand, is the fixing of a variable that may or may not render that variable independent of its non-descendants. I introduce this distinction because, although an ideal experiment will comprise one or more ideal interventions, a single ideal intervention may in turn comprise one or more manipulations. Manipulations are changes to a variable introduced to achieve an intervention. In the ideal case, an intervention into $Z$ will comprise precisely one manipulation, a manipulation to bring $Z$ to the desired experimental level. But there is no reason why, in a single intervention, we must limit ourselves to a single manipulation. Call an intervention in one manipulation 'simple', and in multiple interventions 'complex'.

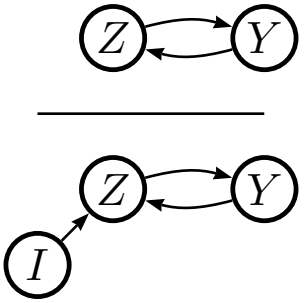


Figure 8.1: $Z$ and $Y$ in a feedback cycle that cannot be broken.

Now, suppose that we are investigating a mechanism as the one in Figure 8.1 in which $Z$ and $Y$ are tied together in a feedback loop. Suppose moreover that a simple intervention into $Z$ is insufficient to cut the causal arrow from $Y$ to set($Z$), in that

[2]Or a set of variables; what I say here applies *mutatis mutandis* to sets.

the act of setting $Z$ changes $Y$, which is itself not a problem, but the change in $Y$ introduces a change in $Z$, causing $Z$ to deviate from the value set by our intervention. The resulting violation of modularity is subtle.

**(PMa)** requires that, if $Z$ does not cause $Y$, then set($Z$) is independent of $Y$. We observe that set($Z$) *is* dependent on $Y$, and so that $Z$ does cause $Y$. But because of the unbroken feedback, and the actual value of $Z$ gets away from the value of set($Z$), we cannot say that the correlation we do observe is the result of our intervention, that is, we cannot report whether we are observing that set($Z$) or just $Z$ *simpliciter* is correlated with $Y$. For example, it might be that, under the experimental circumstances, although $Z$ causes $Y$ generally, here manipulating $Z$ yields no change in $Y$. Yet, it might also be that $Y$ varies randomly, and changes in $Y$ in this context *do* bring about changes in the value of $Z$. Thus, we cannot say that the observed correlation was the result of the setting of $Z$ to a particular value, or the result of random fluctuations in $Y$ driving $Z$. Thus, although application of **(PMa)** yields the right results, it is (possibly) for the wrong reasons. All because the simple intervention fails to break the causal arrow that extends from $Y$ to $Z$. Modularity is therefore not satisfied if we cannot eliminate the causal influence from $Y$ to set($Z$).

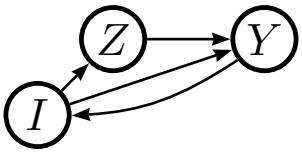


Figure 8.2: $Z$ and $Y$ in a feedback cycle that canceled out by the intervention.

But, if we cannot cut off $Y$'s causal influence on $Z$, we might still achieve the desired result through the simultaneous manipulation of both $Z$ and $Y$. Here is how. We begin by measuring $Y$. Then, we manipulate $Z$ as before, setting it to the desired

command value. At the same time, we also manipulate $Y$ via some distinct route or means, setting $Y$ so as to prevent it from deviating from the measured value. Now, too, we measure how much change we must introduce into $Y$ to hold it at a fixed value—the change in $Z$ changes $Y$, and our second manipulation changes it back. Now, because $Y$ doesn't change, we can vary $Z$ at will. And because we are measuring the manipulation of $Y$, we can use this as an indirect measure of $Z$'s effect on $Y$ independent of $Y$'s effect on $Z$. In this way, we can pull the dependence between set($Z$) and $Y$ apart. The structure of this complex intervention is diagramed in Figure 8.2. Abstractly speaking, the two manipulations of $Y$ and $Z$ together constitute a complex intervention into $Z$ that creates modular independence, in that they permit us to take $Z$ through its full range independently of the value of $Y$, where we could not achieve this independence with just one manipulation of $Z$ alone.

This abstract case may sound like armchair speculation, but in fact this technique of using multiple manipulations in one intervention to induce modularity in an otherwise non-modular system is common. Frequently, researchers are faced with a positive feedback loop linking two variables so as to render them dependent; and to intervene, the loop must be broken. Positive feedback loops can often be broken by constructing an additional negative feedback loop, such as described above, to counteract the positive feedback loop. Indeed, this is precisely how Kenneth Cole's voltage clamp permits the investigation of the voltage-current feedback loops that drive the action potential in neurons.

## The Voltage Clamp

Neurons conduct electrical signals down the length of their axons, called action potentials. The action potential is a moving wave-front that is triggered in the axon hillock, and travels down the axon where it stimulates the release of neurotransmitters. The

action potential works in part by exploiting Ohm's law to generate a positive feedback mechanism.

Consider a segment of the axon. Prior to an action potential, there is a greater concentration of $Na^+$ ions outside the cell than inside. An approaching action potential wave-front drives the membrane potential—the difference in electrical voltage across the axon membrane—higher. Embedded in the membrane are voltage-sensitive gates. The change in membrane potential triggers the $Na^+$ gates to open. When they do, $Na^+$ ions flood into the cell, driven by diffusion forces (against the membrane potential). The rapid influx of $Na^+$ drives the membrane potential even higher, very rapidly. As the membrane potential increases, more and more $Na^+$ channels open, admitting more and more $Na^+$ ions in a positive feedback loop. At this point, the action potential wave-front (which is basically a region of high membrane potential) is propagated into the next adjacent section of the axon. Once the membrane potential reaches a particular threshold level, $K^+$ channels in the cell membrane begin to open, beginning a negative feedback cycle. The opening of the $K^+$ channels permits $K^+$ ions to flow out of the cell, driven by the membrane potential (and against diffusion forces). The movement of $K^+$ ions out the cell rapidly brings the membrane potential down, well below the resting potential, and thus closes the $Na^+$ ion channels and $K^+$ channels as well. Finally, specialized proteins called $Na^+/K+$ pumps actively pump $K^+$ into the cell and $Na^+$ out of the cell, restoring the resting state and readying the axon segment for another action potential.

In general, the action potential is characterized by by two feedback loops: a positive feedback loop linking the membrane potential with $Na^+$ currents, and a negative feedback loop linking the membrane potential with $K^+$ currents. Both of these ionic currents are electrical currents (of a piece with electrical currents generated by the movement of $e^-$ particles—electrons).

The first step to discovering the mechanism for the action potential came with Hodgkin and Huxley's voltage clamp experiments (Hodgkin, Huxley, & Katz, 1952; Hodgkin & Huxley, 1952a,b,c,d). Hodgkin and Huxley had several hypotheses about the mechanism for the action potential (Huxley, 2002), and each hypothesis made different predictions about the specific relationship between membrane potential and each ion current across the membrane. Thus, their first task was to measure this relationship.

Several issues arise with the measurement of ion current at given command levels for the membrane potential (Cole & Moore, 1960; Hodgkin, Huxley, & Katz, 1952; Koester & Siegelbaum, 2000). The first issue is the membrane capacitance. Capacitors are electronic components that pass a current when and only when there is a change in voltage across the capacitor. Because the cell membrane acts as a capacitor, electrically separating the cell interior from the exterior, any change in membrane potential will cause an electrical current to flow across the membrane. This capacitive current, know to Hodgkin and Huxley, is distinct from the currents that result from the working of the action potential, they needed to take independent measurements of the capacitive current so that they could subtract it from their later current measurements.

The second issue is that the membrane potential is not stable. Because the squid giant axon—their preferred preparation—is quite small (no more than maybe one millimeter in diameter), the pool of ions within the axon is likewise relatively small in comparison to the effectively infinite reservoirs of ions outside of the cell membrane. The movement of ions across the membrane in response to a membrane potential thus significantly alters the size of this ion reservoir, either fortifying or depleting it, and therefore alters the relative charge density across the membrane. So, ionic current across the membrane is a cause of the membrane potential. But because voltage

is an electro-motive force, specific values of the membrane potential will drive ions (and electrons) across the cell membrane. So the membrane potential is a cause of transmembrane ionic current. Above a certain potential, ionic current and membrane potential form a positive feedback loop, each driving the other.[3] Thus, any attempt to explore the voltage-current relationship by simply creating a given membrane potential will lead to instabilities in the membrane potential that result in unreliable measurements of current—one cannot be sure how much of the current is due to the command potential, and how much is due to the instabilities in the membrane potential. This second issue is particularly vexing because it exacerbates the first issue: Setting the membrane potential will cause, via feedback, additional changes in voltage which, per the first issue, create a confounding capacitance current. Moreover, this second issue is particularly interesting because it means that simple interventions into the membrane potential cannot possibly satisfy modularity—set($V$) remains dependent on $I$, making a determination of the effect of set($V$) on $I$ impossible.

But both issues can be solved simultaneously. The solution to the second issue lies in finding a way to stabilize the membrane potential at a given command potential, by neutralizing the positive feedback loop. And, since the capacity current is a function of *changes* in membrane potential, and since solving the first issue requires stabilizing the membrane potential, solving the second issue automatically solves the first. But, how to stabilize the membrane potential?

Cole (1949) invented a device that monitors changes in the membrane potential and adjusts the applied voltage via a calibrated negative feedback loop. Cole observed that injecting an *electric* current (*i.e.* a current composed of moving electrons) that opposed the ionic current can be used to fix the total relative charge density—the

[3]This discussion only considers the effects of $NA^+$ ions; $K^+$ ions are also important for the action potential, because their movement across the membrane is part of a negative feedback loop that returns the axon to a resting state.

determiner of the membrane potential—, even while ions are flowing freely across a membrane. If we continuously measure the membrane potential, and continuously adjust the injected electric current to be just strong enough to exactly counter the ionic current, we can keep the membrane potential at the command level. Because the electric current will be exactly opposed to the ionic current, we can measure the electric current being pumped across the cell membrane as a proxy measure for the ionic current. This technique, called the *voltage clamp*[4], holds the membrane potential steady by using negative feedback to counteract the positive feedback cycle described above. The voltage clamp permits interventions into the membrane potential by manipulating first the membrane potential directly, then by manipulating the total transmembrane current by injecting an electric current such that the total current remains at zero (*i.e.* so that ionic current + electric current = 0). The voltage clamp is thus a technique for complex interventions that insure modularity. Here is how the apparatus achieves these ends.

The voltage clamp begins with four wires (Figure 8.3). Two wires, the current wire ($a$) and the voltage wire ($b$), are introduced inside the axon. Two more wires, the reference wire ($c$) and the ground wire ($e$) are introduced just outside the axon.[5] The current wire ($a$) is connected to both an ammeter (not pictured) and the output of a feedback-generating amplifier (whose operation is described below). The voltage wire ($b$) and the reference wire ($c$) are connected to the input of the amplifier via a special kind of voltmeter called a voltage comparator.

The membrane potential is measured by the voltage comparator. One probe—the voltage wire ($b$)—is inserted inside the axon, and the other—the reference wire ($c$)—is placed just outside the axon. The voltage comparator measure the potential between

[4]Cole's name for the device was the 'potential control'; In fact, Hodgkin, Huxley, & Katz (1952) were the first to call it a 'voltage clamp'. (Cole & Moore, 1960)

[5]I use the same wire labels as Hodgkin, Huxley, & Katz (1952); wire $d$, absent from this discussion, is part of a mechanism to verify that the voltage clamp is operating as intended.

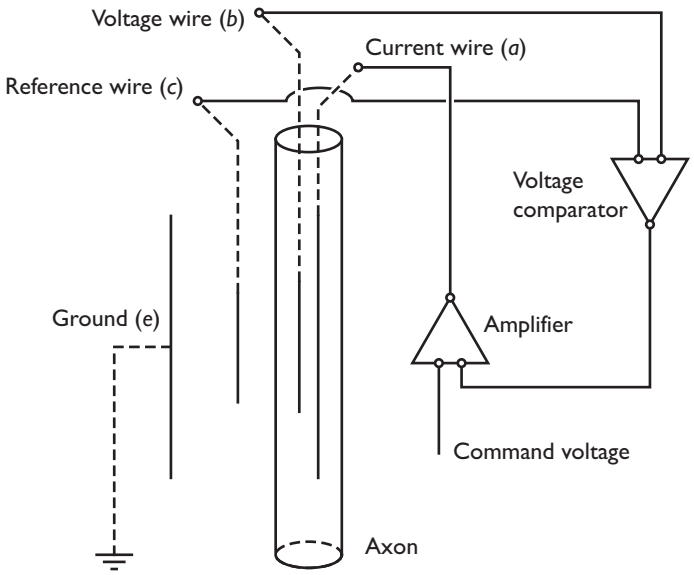


Figure 8.3: Configuration of electrode wires and voltage clamp.
(Adapted from Hodgkin, Huxley, & Katz 1952.)

these probes; and since they are placed on either side of the axon membrane, this potential is the membrane potential. The voltage comparator, on the basis of this measurement, sends a signal to an amplifier. This amplifier compares the signal coming from the voltage comparator to a *command voltage* created by a pulse generator (not pictured). When the two signals differ, the amplifier creates an electric current across the membrane from the current wire ($a$) to the ground ($e$), in this way adjusting the membrane potential.

The electric current generated by the amplifier, in order to successfully stabilize the membrane potential, must exactly match the ionic current across the membrane (since the membrane potential is being held fixed, the capacity current is therefore zero). In this way, even as the ion concentration gradient changes in response to ion flux across the membrane, the total flow of current (both ionic and electric) does not change, and hence neither does the membrane potential. Very conveniently, because the ionic current and electric current are exactly counter-balanced, to measure total

ionic current Hodgkin and Huxley needed only to place an ammeter (again, not pictured) along the current wire ($a$) to measure the electric current; reversing the sign yields the ionic current.

In this way, a command voltage can be set, held steady for a period of time, and ionic current measurements can be read directly off the current wire ($a$), thus permitting the voltage-current measurements needed to begin exploration of the mechanism for the action potential. So the voltage clamp is a technique for intervening via multiple simultaneous manipulations in order to break apart otherwise dependent components. The voltage clamp is an intervention technique that renders the membrane potential independent of its causes, namely the effects of changing transmembrane current. The voltage clamp secures modularity in a system that does not offer independent manipulability of its components.

There is yet another way to secure modularity: The construction of an experimental analog. I turn now to consider this method.

## 8.2 Analog Construction

Sometimes, modular independence fails because the necessary interventions are simply not available for reasons of ethics, logistics, or ignorance. In these cases, there is a second way to achieve as-if modular independence: By constructing an analogous mechanism which does exhibit the necessary independencies, ensuring modularity.

When a mechanism has components $Z$ and $Y$ that are dependent in a way that defeats modularity, we can sometimes design an intervention into $Z$ that satisfies modularity by constructing a second mechanism analogous to the first, but which does exhibit modular independence. Call the mechanism of interest the *target* mechanism, and the constructed analog the *experimental analog* mechanism. The intervention is

then carried out in the experimental analog. If the analog is constructed such that set($Z$) is conditionally independent of $Y$ given $Y$'s parents, then even if interventions into the target mechanism are impossible without violating modularity, the analog still can. And if the analog is sufficiently similar to the target, then the experimental results from the analog can be used to draw conclusions about the target.

Using the strategy of analog construction requires some additional inferential assumptions. First, the analog must differ from the target—otherwise there would be no point in constructing the analog. In particular, the point of difference must be that where set($Z_{\text{target}}$) is dependent on $Y_{\text{target}}$, in the analog set($Z_{\text{analog}}$) *must* be unconditionally independent of $Y_{\text{analog}}$. Often times, establishing such independence is non-trivial: For example, in cognitive psychology, although a particular brain function might exhibit modular independence in that we can lesion brain regions to (literally) cut causal connections, in healthy human subjects such lesioning is staggeringly unethical. Yet, we can still establish as-if modular independence by constructing experimental analogs from elaborate computer simulations, *e.g.* ACT-R (Byrne, Anderson *et al.*, 2004), and simulating lesioning in the model instead. Computer models are also popular for testing how changes to environmental factors will affect local animal populations where directly manipulating the environment is practically impossible (*e.g.*, the predator-prey models of Lotka, 1925; Volterra, 1926).

Second, the analog must not differ from the target *too much*. In general, analogical reasoning stands or falls on the similarity between the target and the analog. It be a mistake to claim that a pocket-watch is a suitable experimental analog for the suprachiasmatic nuclei in mammalian brains, simply because both exhibit a 24-hour cycle. The analog must be sufficiently similar to the target to support analogical inference from interventions on the experimental analog to causal conclusions about the target. Thus, we must already have some basic knowledge of the mechanism in

order to take advantage of this technique.

Thus, there are a matched pair of constraints on analog construction. Here is one articulation of these constraints that I will defend below. Suppose that we have a mechanism with components $Z_{\text{target}}$ and $Y_{\text{target}}$, in which $Z_{\text{target}}$ does not cause $Y_{\text{target}}$, and in which we desire to intervene on $Z_{\text{target}}$, and where set($Z_{\text{target}}$) is dependent on $Y_{\text{target}}$. Then, two constraints that are jointly sufficient for an inference from the analog to the target are:

**Causal Difference** The analog must differ from the target insofar as $Z_{\text{analog}}$[6] must not depend on any other components besides its (direct and indirect) effects. That is, $Z_{\text{analog}}$ must have no causes (including latent common causes connecting $Z_{\text{analog}}$ to any other measured variable), and must not be otherwise correlated with any other components (except its effects).

**Causal Similarity** Aside from the constraints on $Z_{\text{analog}}$, the causal structures of the two mechanisms, including the functional nature of those causal relations, must be the same across the target and the analog with respect to the set of measured variables. That is, $\forall X \forall V \neq Z$ if $X_{\text{target}} \rightarrow V_{target}$ then $X_{\text{analog}} \rightarrow V_{\text{analog}}$ and if $V_{target} = f(X_{target})$ then $V_{analog} = f(X_{analog})$.

The first constraint specifies how the analog must *differ* from the target: If we are intervening into $Z$, then we must construct our analog so that $Z_{\text{analog}}$ has the independence necessary to permit that intervention without violating modularity. Recall that the presumption is that such an intervention is not possible in the target mechanism (else we wouldn't need to construct an analog). So, we must be sure that our construction renders $Z_{\text{analog}}$ exogenous, where such is not true about $Z_{\text{target}}$. By rendering $Z_{\text{analog}}$ exogenous, we can be certain (if our intervention is ideal) that we can

[6] Or, more precisely, whatever it is that $Z_{\text{analog}}$ represents in the analog mechanism.

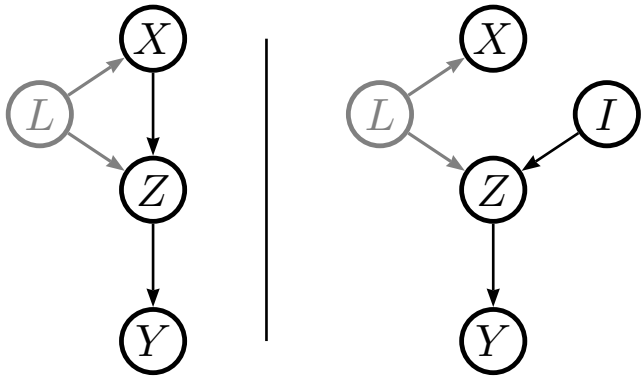


Figure 8.4: The problem of latent common causes. Left: Target. Right: Analog

manipulate $Z_{\text{analog}}$ secure in the knowledge that any correlation that arises elsewhere in the analog arises as an effect of $Z_{\text{analog}}$.

Notice that it is not enough to cut the causal arrows from *measured* variables to $Z_{\text{analog}}$. Although we do not have to cut *every* cause of $Z_{\text{analog}}$ off, we must at least cut off the influence on $Z_{\text{analog}}$ from latent common causes that connect $Z_{\text{analog}}$ to other measured variables, because such common causes can create confounding dependencies in the experimental analog. Consider the case illustrated in Figure 8.4. On the left is the target mechanism, in which $X_{target}$ causes $Z_{target}$; $Z_{target}$ causes $Y_{target}$. But, unknown to us, there is a latent variable $L_{target}$ that is a common cause of $X_{target}$ and $Z_{target}$. In the analog, we must cut the arrows from each measured variable that is a cause of $Z_{analog}$; ($I$ represents our intervening into $Z_{analog}$). But our intervention will still fail modularity, because we will find that $X_{analog}$ remains correlated with our intervention into $Z_{analog}$ even though $Z_{analog}$ does not cause $X_{analog}$. So we would—falsely!—conclude that $Z_{target}$ was a cause of $X_{target}$. The reason is that the latent common cause $L_{analog}$ creates a confounding dependency between $X_{analog}$ and $Z_{analog}$. So we must be sure that our analog cuts arrows to $Z_{analog}$ from latent common causes.

The second constraint specifies how the analog must remain *similar* to the target.

Without similarity, there is no basis for analogical inference from the model to the target. Since the reasoning being deployed is causal, this constraint specifies causal and functional similarity as the relevant similarity metrics. Thus, although the target and the analog may be composed of different components in different materials, the functional and causal relations must remain the same (excluding the restrictions on $Z_{analog}$ specified in the previous constraint). That is, the set of measured variables which we have picked to represent the target must be able to represent components in the analog, such that the causal structure among the variables is the same for both the target and the analog.

It is important to emphasize that similarity of bare causal structure is not enough: We must ensure the functional relations are the same as well. Suppose that the target mechanism consists in two cogs, one twice as large as the second, and that the larger drives the smaller. In our model, $A_{target}$ represents the rotation of the larger, and $B_{target}$ the rotation of the smaller. Thus, in our model, $A_{target} \rightarrow B_{target}$. Now, consider a second mechanism that consists in two levers, one $A_{analog}$ half as large as the second $B_{analog}$, and that the smaller lever drives the larger. Thus, in this second model, $A_{analog} \rightarrow B_{analog}$. The causal structures of the two models are identical, but the functional relation differ across the two models, because $B_{target} = \frac{1}{2}A_{target}$, yet $B_{analog} = 2A_{analog}$. The functional relation between $A$ and $B$ is different across the two mechanisms. And so, interventions into lever $A_{analog}$ with respect to lever $B_{analog}$ will *not* yield reliable information about the functional relationship between the two cogs in the first mechanism. For this reason, experimental analogs must not only share the same causal structure as the target (with respect to the measured variables), but the same functional structure as well. But just as the entities need not be the same, neither need the activities be the same. What is important for analogical inference in this case is that the mathematical relationship between the

components (the active entities) remain the same. Whether we know in advance that the mathematical relationships are the same is not itself important; thus, we can use this method to learn about a causal system, verifying the necessary verisimilitudes *post hoc*.

Now, because I have not specified any constraints on the particular entities and activities used in constructing an experimental analog, we can be quite creative in how we satisfy the first constraint. One way to intervene into Cartwright's tightly-coupled carburetor, for instance, is to simply replace the Venturi chamber with an array of new components, each of which plays one and only one of the several roles the Venturi chamber plays.[7] So long as each of the new parts plays the same causal and functional roles in the analog as the Venturi does in the target, then we can treat the analog as equivalent to the target with respect to interventions into any of those new components. That said, this strategy clearly requires some causal knowledge of the mechanism in order to assess the quality of the original; but we do not need a complete knowledge of the causal or functional structure (otherwise there would be no point in experimenting in the first place), as the example below will demonstrate.

We can see these constraints at work in actual scientific practice. Experimental analogs may be the most common technique deployed by scientists generally. Nearly any experiment that requires an experimental preparation, or that makes use of a model organism, or that uses a computer or physical simulation makes use of this strategy. Consider the intriguing case of Loewi's discovery of *vagusstoff*, a chemical we now know as acetylcholine.

[7]Which, incidentally, is precisely what a fuel injector does—it performs the same functions as the carburetor, but in such a way that all of the various functions of the Venturi chamber in a carburetor are carried out by separate components.

## Loewi's Beating Hearts

Loewi's discovery of *vagusstoff* (Loewi & Navratil, 1926) is a clear example of creating as-if modular independence using analog construction. Acetylcholine, we now know, is the medium of communication between the vagus nerve and the heart. The vagus nerve is one of two major nerves connected to the heart; the vagus nerve signals the heart to slow its rate of beating (and the other nerve signals the heart to increase its rate of beating). In the late nineteenth and early twentieth centuries, biologists sought to understand the mechanism by which these nerves were capable of exciting or inhibiting the muscle fibres of the heart. Although Galvani theorized that the vagus nerve used electrical signals to directly stimulate the heart muscles, by the mid-1920's there was good reason to think that the vagus nerve might instead rely on chemical signals.

In 1921, Loewi performed a now classic experiment to decide whether the vagus nerve was signaling the heart via electrical impulses or via chemical secretions (Loewi & Navratil, 1926). Loewi prepared two frog hearts by cannalizing them and submerging them in Ringer's solution. In the first heart, he electrically stimulated the vagus nerve. If the vagus nerve communicated with the heart muscles by secreting a chemical agent (which, not knowing what that chemical would be, or whether it even existed, he referred to with the delightfully vague term '*vagusstoff*'), then the vagus nerve should, he reasoned, secrete that agent into the Ringer's solution. He then collected a sample of the Ringer's solution from and transferred it to a second heart preparation, which differed only in having been stripped of its vagus and sympathetic nerves. If the transplanted Ringer's solution contained *vagusstoff*, then the muscles of the second heart should respond to it by slowing its beating.

Indeed, the second heart did slow at the introduction of the Ringer's solution from the first heart. Although it would be several years before *vagusstoff* was shown to be

acetylcholine, Loewi's demonstration was sufficient to establish that the vagus nerves do signal the heart via a chemical agent. Let us consider his demonstration in the light of my discussion of analog construction.

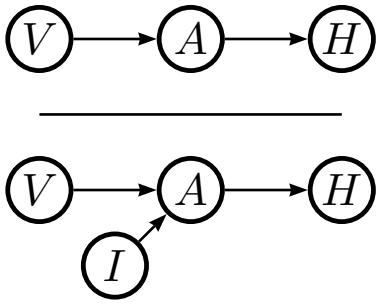


Figure 8.5: Causal models of the mechanism for the vagus nerve, and a one-preparation experiment that violates modularity. Although in this experiment, $V$ is what is actually being manipulated, the intended intervention is in fact on $A$—hence the violation.

Loewi had clearly defined input and output components, the excitation of the vagus nerve and the slowing heart-rate. And, stimulation of the vagus nerve clearly produced a slowed heart-rate, so the input component is a cause of the output component. Loewi hypothesized that there was a very simple mechanism that linked the input and output components, namely *vagusstoff*, as in Figure 8.5. However, not knowing what *vagusstoff* was, or how it acted, there was no way for him to manipulate this component *in situ*, with just one preparation. Since, if *vagusstoff* was real, the only method that Loewi had for producing *vagusstoff* was to stimulate the vagus nerve. So, a one-preparation experiment would violate modularity, because interventions into the presence of *vagusstoff* could not be made independent of the direct cause of *vagusstoff*, namely the the vagus nerve. Figure 8.5 models a one-preparation experiment; $V$ represents the stimulation of the vagus nerve, $A$ the production of *vagusstoff* (acetylcholine), and $H$ the slowing of the heart-rate. In a one-preparation

experiment, $V \rightarrow A$, but set$(A) \not\perp V$ (because there is not way to cut the causal connection linking $V$ with $A$), in violation of **(PMa)**.

To ensure his inferences were sound, Loewi constructed an analog: a second heart preparation *sans* vagus nerve. In this way, his intervention would satisfy modularity, because his intervention into the presence of *vagusstoff* could be made independently of any stimulation of the vagus nerve—for there was no vagus nerve to stimulate. In the analog, he literally cut the causal connection between $V$ and $A$. Thus, any observed changes in the heart could be attributed solely to the intervention—to the presence of *vagusstoff*—, and not to a confound caused by a violation of modularity. To obtain the hypothetical *vagusstoff*, Loewi stimulated the vagus nerve in the first heart; he needed only to introduce it to the second preparation to test his hypothesis. And if, as actually occurred, the introduction slowed the heart rate, then he could infer that the slowed heart rate in the first preparation resulted from the secretion of *vagusstoff* during the stimulation of the vagus nerve: His intervention into the analog provided sufficient grounds for reasoning about the causal structure of the original preparation.

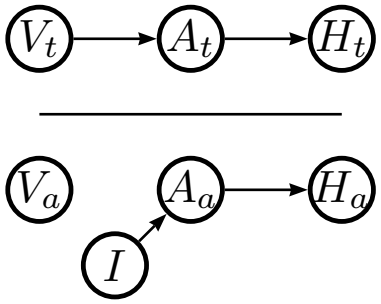


Figure 8.6: Causal models of Loewi's two-preparation experiment. Top, the *target*; bottom, the *analog*.

Figure 8.6 presents a model model of Loewi's two-preparation experiment. At the

top is a model of the first preparation, the target mechanism. At the bottom is a model of the second preparation, the experimental analog. Loewi could not intervene into $A_{target}$ without also necessarily intervening into $V_{target}$ (because the vagus nerve was the only means of producing *vagusstoff* that Loewi knew of, not knowing what *vagusstoff* was or even if it existed); so he created an analog in which he could cut (quite literally) the causal influence of the vagus nerve on the production of *vagusstoff*, and so intervene directly into $A_{analog}$.

Finally, having established that *vagusstoff* was capable of slowing the heart-rate in the experimental analog, Loewi reasoned that the that *vagusstoff* was the intermediate mechanism by which the vagus nerve slowed the heart-rate in the target mechanism, the first preparation.

There are three points to make about this example. First and foremost is that the target preparation violated modularity (and not for any metaphysical reason, but because Loewi simply didn't know enough about his hypothesized *vagusstoff*—even whether it existed—to synthesize it by means other than via the vagus nerve itself), so that Loewi couldn't safely assume that the components exhibited modular independence. Second, Loewi overcame this hurdle through the construction of an analog mechanism that permitted him to reason as if the components did exhibit modular independence anyway. Third, Loewi then used a form of analogical reasoning to apply the experimental correlation he observed in the analog preparation to the mechanism at work in the target preparation.

We can see why the arguments against modularity glossed earlier are unconvincing. The critics' arguments from modular independence presume that experimental interventions take place within the very same mechanism we are trying to understand. This may be true of some kinds of mechanisms (especially those that really do, for example, provide affordances), but it is not true generally, and certainly wasn't

true for Loewi's case. Since modularity is a modal claim about what would happen during an intervention, modularity need only hold in the experimental analog and not of the target system. Thus, there is no particular reason to think that seemingly non-modular systems, such as Leowi's first preparation, or carburetors and toasters, violate modularity *full stop*. Each of these mechanisms can be investigated by testing causal claims in an experimental analog.

But, as I've already said, we are not free to construct our analog in any way we wish. Recall the two constraints I raised earlier. The first constraint requires that, if we are intervening into $Z_{analog}$, that $Z_{analog}$ be uncorrelated with its non-effects in the analog. Loewi's second preparation eliminated the component that produced *vagusstoff*, the vagus nerve. Doing so allowed him to introduce *vagusstoff* into the preparation without worry about this kind of confound.

The second constraint requires that the causal and functional structure of the analog mirror the target as closely as possible (ideally, exactly). Leowi ensured that his analog preparation had the same causal structure as the target (aside from the severed vagus nerve), by using the same kind of heart and the same kind of preparation as in the target. Presuming that the neither heart was in any way abnormal, by constructing the analog out of the same materials as the target, Loewi was able to readily satisfy the second constraint. I should take pain to note, however, that he was in no way *constrained* to use a heart from the same species of frog; rather, he used a heart from the same species of frog because it was a very expedient and uncontroversial way to ensure that both preparations shared the same causal structure.

Having ensured his analog satisfied these constraints, Loewi was therefore justified in using modularity to infer from his experimental observation of the analog preparation—that the introduction of Ringer's solution slowed the heart rate—to a conclusion about the causal structure of the target preparation—that the vagus

nerve communicates with the heart via a chemical soluble in Ringer's solution. What licensed this inference was that *vagusstoff* has the same causal and functional influence on the analog heart as on the target heart—and *not* that both hearts were made of the same stuff.

(But notice too, as an aside, that Loewi was not free to use, for example, a contrived mechanical heart designed specifically to respond to, say, Ringer's solution. Precisely because such an analog is contrived, and because Loewi *didn't* know the mechanism by which the vagus nerve worked, we wouldn't have sufficient reason to think that such a 'heart' would have shared the same causal and functional structure as the target heart. Thus, Loewi would not have been justified in drawing inferences from the contrived 'heart' to the target frog heart.)

I turn now to demonstrate a third way that as-if modular independence can be created: the method of subtraction. The method of subtraction is effectively a special case of analog construction, but sees widespread enough use to warrant a separate discussion.

## 8.3 The Method of Subtraction

There are many occasions when neither multiple interventions nor analog construction can be used to effectively (or ethically) ensure modularity. The problem is that oftentimes we cannot independently manipulated the components in a mechanism, not because of anything about the mechanism—it might well be modularly independent—but because we find ourselves limited to ham-fisted interventions. Many human studies suffer this difficulty: There is no currently known way to conduct a drug study, for example, that does not risk confounding the effects of the drug with the placebo response. Clinical pharmacologists rely on a method for studying analgesics that uses a

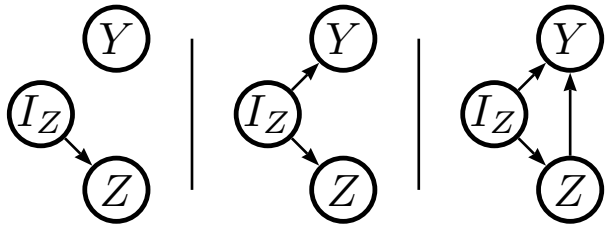


Figure 8.7: Three different interventions; Left: surgical intervention into $Z$; Middle: Ham-fisted intervention into $Z$; Right: Ham-fisted intervention into $Z$ that obscures the causal relation between $Z$ and $Y$.

second intervention (called a *control*) that differs from the experimental intervention only in that the drug administered is a sham. In the control patients, any response to the inert pill administered can be reasonably attributed to the placebo response. This placebo response is measured in the control case, and then *subtracted* from the measured response in the experimental case. In this way, researchers can draw inferences as though the responses to the drug were independent from the placebo response.[8] Mill (1843) calls such a procedure the *method of difference*.

One way that modularity can be violated is when our interventions into $Z$ necessarily, though unintentionally, manipulate $Y$ (whatever the causal relationship between $Z$ and $Y$ might be). Figure 8.7 illustrates the point: On the left, the intervention affects only $Z$—call such interventions that affect only the targeted components *surgical*. In the middle, the intervention affects $Z$ *and* $Y$, although it was only targeted at $Z$—call such interventions *ham-fisted*. Such interventions violate modularity when set($Z$) would have been independent of $Y$ had the intervention been surgical, but the ham-fisted nature of the actual intervention leads to a dependency between set($Z$) and $Y$. Moreover, even when $Y$ *is* an effect of $Z$, as in the right of the figure, the functional nature of that dependency will change as a result of the intervention, and

[8]Notice that the method of subtraction can be used in conjunction with multiple interventions and analog construction, and is not an exclusive alternative to those methods. For example, in analog construction, control analogs are often required to verify the similarity of the analogs to the target.

even though we would correctly observe that $Z$ was a cause of $Y$, we could yet be quite mistaken about the functional relationship between $Z$ and $Y$.

Notice, too, that although in the first two cases, the surgical and the ham-fisted interventions are both technically ideal, in the third case, the ham-fisted intervention is *not* ideal. So, if we are presuming ideal interventions, the third case should never arise. In which case the issue at hand is not that the mechanism fails to exhibit modular independence, but that we cannot achieve an ideal intervention. Some non-ideal interventions, such as a ham-fisted intervention, will create a violation of modularity in an otherwise perfectly modular mechanism.

When ham-fisted interventions are unavoidable—as for example in pharmacological studies—, there is yet a way to satisfy as-if modularity. We can do so by making *two* interventions[9]—one *experimental* intervention and one *control* intervention[10]—into *two* subjects—the experimental subject, and the control subject. In the experimental case, we simply perform the full-blown ham-fisted intervention. In the control case we perform a limited intervention that manipulates only the ham-fisted bits of the experimental intervention, and not the intended target of the experimental intervention. We then measure the changes arising from the control intervention, and record them as error. Finally, we subtract (often literally, but sometimes metaphorically) the error from the results of the experimental case. Ideally, we are left with a measure of the changes that would have been brought about had the experimental intervention been surgical.

The particulars of the subtraction will vary depending on the particulars of the experiment. In neuroimaging studies, for example, two images—one experimental and one control—are overlaid, normalized, and registered; whatever is common to both images is subtracted out with a linear filter. In response time studies, subjects are

[9]Not *manipulations*, but full *interventions*.

[10]Thus, we may need two sets of multiple manipulations or two analogs to make two interventions.

timed performing a baseline task and an experimental task, and the response times are numerically subtracted. The results of each subtraction are taken as a measurement of the direct effects of the intervention on the dependent variable.

There are, of course, constraints on how the control intervention should be crafted. Suppose that we have a set of variables $\mathbf{V}$, and that we wish to intervene into $Z \subset \mathbf{V}$. But suppose that our intervention is ham-fisted, and also affects all of the variables in some distinct subset $\mathbf{H} \subset (\mathbf{V} \setminus Z)$. To justify the subtraction of the difference between the control and the subject, and hence create as-if modular independence, a control intervention must satisfy the following conditions:

**Subject Similarity** The control subject and experimental subject must share the same causal and functional structure with respect to $\mathbf{V}$;

**Control Similarity** the control intervention must manipulate all and only the members of $\mathbf{H}$ in the control subject, and must do so in exactly the same way as the ham-fisted experimental intervention in the experimental subject; and

**Control Difference** the control intervention must *not* manipulate $Z$ in the control subject in any way.

That is, we must be certain that our control intervention is ham-fisted in precisely the same ways as our experimental intervention; only then can we be sure we are subtracting out (controlling for) all and only the ham-fisted aspects of the experimental intervention.

These constraints are illustrated in the four diagrams of Figure 8.8. Suppose that we wish to intervene into $Z$, but we can only do so ham-fistedly, as in the first diagram, such that $\mathbf{H}$ contains $W$ and $X$. The second diagram illustrates how our control intervention *should* be structured so that we can apply the method of subtraction. The third and fourth diagrams show two different ways that the control interventions

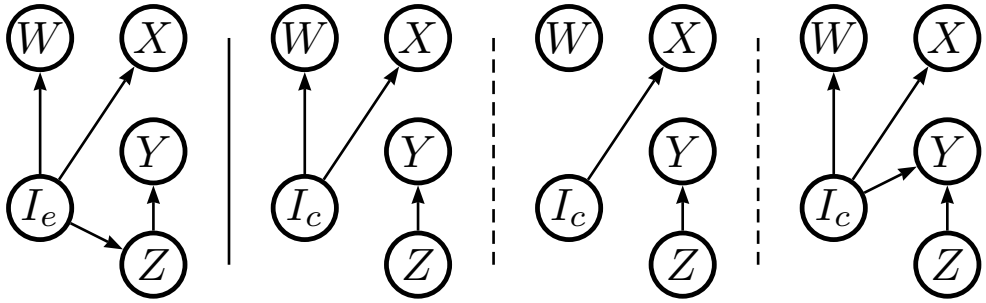


Figure 8.8: A controlled intervention; Left: Experimental, ham-fisted intervention; Middle-Left: Control intervention; Middle-Right and Right: Botched control interventions.

can go wrong: They can control too little, or too much (or both, not pictured, by affecting only $X$ and $Y$).

If the control intervention does not affect *all* of **H**, as in the third diagram in Figure 8.8, then we will be led to draw false positive conclusions that $Z$ is a cause of whatever variables in **H** are left out of the control, and which co-vary with set($Z$). In the diagram, for example, our control intervention does not control for $W$, and we will be falsely led to conclude that $Z$ causes $W$.

On the other hand, if the control intervention affects more than what is in **H**, and those additional variables are effects of $Z$, then we can be falsely led to conclude that $Z$ is not a cause of those variables, because they won't be seen to co-vary with $Z$ after subtraction. In the fourth diagram, for example, our control intervention controls for $Y$ when it shouldn't. As a result, we may be falsely led to conclude that $Z$ does not cause $Y$.[11]

To illustrate the method of subtraction in context, I turn to the study of analgesic effectiveness in humans. Evaluating the effectiveness of pain relievers requires two applications of the method of subtraction, and we can use the forgoing discussion to

[11]That wiggling $Z$ does not yield concomitant wiggling of $Y$ is not conclusive proof that $Z$ does not cause $Y$, but it is at least *prima facie* evidence that $Z$ likely doesn't cause $Y$.

understand the complexity in this standard experimental design.

So, suppose we wish to evaluate the effectiveness of a new analgesic drug for pain relief. Because pain tolerance is highly subjective, we cannot measure an analgesic's effectiveness in absolute terms. Rather, we must measure how much of a *difference* in a subject's pain tolerance the analgesic makes. One way we can measure this difference in tolerance is with the *cold pressor task*, invented by Wolf & Hardy (1941). In this task, the subject is asked to place her hand in a bowl of ice water for as long as she can tolerate it—very painful, but only so long as the hand is submerged. The length of time that she can tolerate leaving her hand submerged is taken as a measure of the subject's base pain tolerance. We might measure an analgesic's effectiveness by administering the drug and asking the subject to perform the cold pressor task. The resulting measurement is alone, however, is insufficient to tell us how effective the drug is, because our intervention was ham-fisted: We have no way to distinguish the effects of the drug from various other factors that yield (operationalized) pain tolerance (*e.g.* that it takes time to cool the hand sufficiently to sense that cooling as pain, that pain signals take non-zero time to travel from the hand, that there is some neural mechanism that suppresses immediate response to pain, *&c.*). Thus this simple intervention results in a violation of modularity.

To measure the effectiveness of a pain reliever, we need to control for the laws of thermodynamics, the finite speed of neural signaling, and the subject's baseline pain tolerance: We must measure their pain tolerance first without the drug, and then second with the drug[12]. In this way, we can measure directly the effects of the drug on pain-tolerance levels by subtracting the first measurement from the second. The difference in times can then be attributed entirely to the drug. If the subject is capable of performing the cold pressor task for a longer period in the presence of the

[12]And, perhaps at several intervals after drug administration, to get a time-profile for the drug's effectiveness.

analgesic than without, we might conclude that the analgesic successfully increased the subject's pain tolerance, and we have an objective measure of the size of the effect in the difference in measurements. We are justified in this conclusion, because the first test—the control intervention—satisfies the criteria above. Subject Similarity is satisfied in virtue of taking the two measurements of the same subject. Control Similarity is satisfied in that we are administering the same test in the same way in both cases. And Control Difference is satisfied because the drug is only administered in the second case. Thus, we can be reasonably sure that the difference in measured times is due exclusively to the administration of the drug.

However, although an improvement, even the addition of this second, control, intervention is insufficient to satisfy modularity because analgesics are *double-acting*: they exert an analgesic effect through both a pharmacological route and through the placebo response. Because the drug can affect pain tolerance along two distinct routes, our intervention remains ham-fisted. In the first way, the drug works through whatever biochemical mechanism the drug was designed to target. In the second way, the drug works through much higher-level psychological mechanisms of subject expectations and conditioning, that is, via a second route that has nothing to do with the actual chemical composition of the drug. So, as they stand, our interventions remain ham-fisted, insofar as they cannot distinguish the effects of the chemical action of the drug and the placebo response.

Clinical trials of analgesics therefore make use of an additional control, that we might subtract the placebo response from the pharmacological effect of the analgesic. Consider a typical example: Yuan, Karrison *et al.* (1998) examined the effect of acetaminophen on cold pressor response by having each subject perform four distinct tasks. In the first task[13] the subject was given a small dose of acetaminophen; in

[13]The order of task presentation was randomized from subject to subject, of course. The role of

the second, a medium dose; in the third, a large dose. In the fourth task the subject was administered a sugar pill. Then, the pain threshold responses for each task were compared. Since neither the subject nor the experimenter knew which drug was being administered in each task, the size of the placebo response should be the same in each task. But the quantity of acetaminophen was varied across the tasks from between none to a large dose. Thus, for any given subject, any *difference* in pain threshold response across the tasks could be attributed to the pharmacological action of the acetaminophen; the response for the placebo-only control task is a pure placebo response that is also taken as a baseline measurement of pain threshold. This baseline can be subtracted from the three experimental conditions to obtain the absolute change in response due to the pharmacological action of the acetaminophen alone.

Subject Similarity is maintained by giving each subject both experimental and control tasks (although there might be differences in the subject between trials, of course). The tasks are separated by sufficient time to prevent the subjects from acclimating to the cold pressor task, and to prevent and cumulative effects of being administered several doses of drug close together in time. Control Difference is secured by the use of the placebo task as a control. The placebo has no pharmacological component, and so cannot be used to manipulated the target mechanism, but can be used to manipulate the undesired, placebo mechanism. Finally, Control Similarity is secured by ensuring that the placebo response will be the same across all four tasks. The technician did not know which pill is the sugar pill (and hence inadvertently tipping off the subject), the sugar pills were the same size, shape, and color as the acetaminophen. Since the placebo response seems to result from the subject's knowledge and expectations, and since their knowledge and expectations are carefully managed to be identical across all four tasks, the assumption of identical placebo responses

randomization is beyond the scope of this chapter, but is yet another way to create as-if modular independence.

appears justified.

This kind of clinical drug trials nicely illustrate what I want to say about the method of subtraction. Often, our interventions are ham-fisted, but where we have a method for varying our interventions (as in the administration of either the drug or a placebo), we can use two variants of the intervention to make two ham-fisted measurements. When the interventions differ in precisely one way—where one affects $Z$ and the other does not—then we can subtract the resulting measurements to get at the difference our intervention made to $Z$'s effects directly, despite the ham-fisted nature of those interventions.

## 8.4 Conclusion

In the last four chapters of this dissertation, I have considered that modularity may well be the manipulated mechanism's weakest link. Modularity certainly has its share of critics. The general worry about modularity seems to be that modularity is not a very good *general* principle, that it is too strong to be useful outside of a small number of overly contrived examples.

In response, I have established an intuitive distinction between the modularity as a feature of a mechanism (modular independence), and modularity as a modal feature of an intervention (as-if modular independence). I have shown that, although modularity is satisfied by mechanisms that exhibit modular independence, modularity can also be satisfied, *contra* the critics of modularity, by interventions that create as-if modular independence. Modularity fails only when our intervention is not independent of its effects. But, as the critics I have addressed have rightly noticed, many of our interventions are *not* independent of their non-effects. Frequently, either because of the structure of the mechanism under investigation, a failing of our intervention

technique, ethical issues, or simple ignorance, we cannot be sure that our intervention is independent in the right ways. Nevertheless, we can still secure as-if modular independence, and hence modularity, in these circumstances.

In this chapter, I have applied this concept of as-if modular independence to three historical cases, to show that this view of modularity permits us to make sense of the complex experimental techniques used in each case—an improvement over earlier conceptions of modularity (such as modularity's critics have in mind). I have used these cases to offer some strategies by which we might use interventions to create as-if modular independence. We can use multiple manipulations to directly control for any undesired effects of our intervention; we can construct an analog which is sufficiently similar to the target system, and which is designed to exhibit modular independence; we can treat undesired effects of our intervention as error, measure that error independently from the intervention, and use the method of subtraction to infer what the effects of an ideal surgical intervention would look like. There are doubtless many more such techniques; I need not document them all here. Rather, I need only establish that modularity is not the overly-strong constraint on mechanisms that its critics have claimed it to be, and that even under this weaker interpretation it is capable of doing the inferential work it was designed for. Better yet, this result is only really news to philosophers: Scientists have been relying on as-if modular independence for centuries. What remained was only to formalize these techniques, and show that they are a natural fit to the formal techniques used in the manipulated mechanism.

I conclude that, in contrast to the opinion of critics, modularity is far from the weakest link in the manipulated mechanism; Indeed, the opposite is true, for I have shown that a careful examination of modularity allows us to view many common experimental paradigms as instances of Woodwardian manipulationist inference, and

that modularity is a robust assumption that yet holds a great deal of inferential power.

Thus, I conclude my defense of the manipulated mechanism against the claim that its reliance on modularity undermines the entire project from the start.

# Conclusion

In this dissertation, I have argued for two broad claims. First, I have argued that my account of the Manipulated Mechanism is the right step forward in resolving for the program of mechanistic explanation pressing issues of normativity. Second, I have argued that, whatever other flaws it might have, modularity is not a point of weakness for my account, but a strength.

## The Manipulated Mechanism

Several normative issues face any account of mechanistic explanation. Mechanism discovery is an inferential process, wherein one infers causal structures from experimental data, and is hence amenable to a normative analysis. We might ask: What are the assumptions that justify this kind of inferences? Other normative questions arise, too (Craver, 2007): When is a component properly said to be part of a mechanism? How can we distinguish proper components from mere background conditions or spurious correlates? How can we know when a mechanistic explanation is a good one, that is, is justified by the empirical evidence? Such questions cry out for the addition of a normative component to our understanding of mechanistic explanation.

Although we already have several richly qualitative accounts of mechanistic explanation, none of them provides the necessary normative framework for answering these kind of questions. I have argued that one promising avenue to explore is Woodward's

manipulationist account of causal explanation (Woodward, 2003).

On Woodward's view, a claim that $X$ causes $Y$ is true just when the counterfactual claim that ideal interventions into $X$ would be followed by regular changes in $Y$ is true. Woodward claims that experimentation is one way to test this counterfactual claim—in effect making the counterfactual factual, and therefore readily testable. Woodward defends this claim by providing a set of constraining principles that, when satisfied, permit the inference from correlation observed during experimental intervention to general causal claim.

Woodward himself (2002), Craver (2007), and most recently Glennan (personal communication), have attempted to construct a bridge between Woodward's system of experimental inference, and the qualitative accounts of mechanistic explanation due to Glennan (1996, 2002) and Machamer, Darden, & Craver (2000). However, these first steps are yet incomplete, as none quite manage to bridge the quantitative to the qualitative: Woodward's account doesn't capture all of the detail of the qualitative accounts; Craver's doesn't quite capture all of the detail of the quantitative.

I have argued that the right way to bring manipulationism and mechanism together in a rapprochement that can harness the full inferential power of manipulationism while doing justice to the richly qualitative aspects of the accounts of MDC and Glennan. Where the causal models in Woodward's manipulationism are, by design, largely uninterpreted, the qualitative accounts of mechanism can be thought of as providing constraints on the possible interpretations of a causal model. Thus, speaking very coarsely, mechanisms are bounded entities, with well-defined start and stop conditions, and are composed of entities and activities. Any causal model of a mechanism must reflect these features: It must reflect that mechanism's boundaries, it's start and stop conditions, and its composition.

I have argued for a set of principled constraints that bridge these qualitative

features of mechanisms with the formal apparatus of manipulationism, which taken together I call the Manipulated Mechanism. First, the causal structure of a mechanism model must meet certain structural constraints. A mechanism must comprise one or more variables that represent the explanandum cause, and one or more variables that represent the explanandum effect. A mechanism model must comprise one or more variables that m-separate the explanandum cause from the explanandum effect. Second, the formal elements of a mechanism model must be capable of representing the mechanism using the semantics of the Interactivity View. The arrows must represent not just causal relations, but relations of dependence under intervention. The variables must represent entity-activity pairs. Only models that meet these constraints can be said to be models of mechanisms, and hence candidates for genuine explanatory texts.

In addition to providing a normative framework for answering questions of explanatory relevance, the Manipulated Mechanism provides tools for the discovery of mechanistic explanations from experimental data. To do so, in the final chapter I turned to three particularly complex historical cases of mechanism discovery, cases that appear to defy straightforward manipulationist explanation. These cases illustrate that the Manipulated Mechanism (when paired with a suitable modularity principle, namely **(PMa)**) can account for the various strategies used by the researchers in these cases.

But the Manipulated Mechanism appeared to remain highly vulnerable to a class of objections to the principle of modularity, a cornerstone of Woodward's manipulationism.

## Modularity

Any appeal to Woodward's manipulationist framework faces a particular difficulty: The assumption of modularity. Central to manipulationism is the idea that it should always be possible in principle to intervene into a causal system surgically, that is, without disrupting anything in the system beyond the experimental target. Yet, many complex causal systems defy such surgical interventions—and some of the most interesting mechanisms in biology appear to be among them. If there are known biological mechanisms that are not modular (and yet were discovered via experimental manipulations), then my account of the Manipulated Mechanism will be incomplete, for it will not be able to account for these mechanisms.

Thus, in defense of the Manipulated Mechanism, I have executed a defense of modularity against these claims. My defense is two-pronged: One the one hand, I claim that some charges of violations of modularity are spurious, either misunderstanding modularity or the nature of the violation. No doubt, the term 'modularity' draws Fodorian images in many readers' heads, and yet, Woodward's modularity is quite different. Thus, for example, the brain can be Woodward-modular without being Fodor-modular—and hence any claim that a system violates Fodor-modularity is not automatically a claim that that system violates Woodward-modularity. Too, some commentators (*e.g.* Chemero & Silberstein, 2008) have claimed a close connection between modularity and the mathematical assumption of additivity, to then argue that non-linear dynamical systems, being non-additive, violate modularity. But the connection is not so close: Although additive systems are modular, I argue, not all modular systems are additive.

On the other hand, some charges that certain systems violate modularity are well-founded. I argue that modularity, as Woodward and others have formulated it, really encapsulates two distinct assumptions: That an intervention is probabilistically

dependent *only* on the intervened-into-variable and its effects **(PMa)**; And that an intervention is probabilistically independent of the intervened-into-variable's effects, conditional on those variable's parents **(PMb)**. The first claim is not terribly contentious, but the second is quite strong and quite controversial. I demonstrate that the systems that do violate modularity do so in virtue of violating the second distinct assumption. Moreover, I argue that the second principle is not doing much philosophical work for the modularity assumption, and that we can therefore safely jettison it from our formulation. What remains is weak enough to dodge the counter-examples raised against modularity, yet strong enough to do modularity's work.

I closed the dissertation (as mentioned above) by showing how **(PMa)**, properly understood, is a modal principle that licenses the analogical reasoning from experimental subjects to their 'real-world' counterparts. I have argued that modularity does not require independently manipulable parts, but rather that we can create modularity, via careful interventional strategies, in systems that critics took to be exemplars of non-modularity. Looking at case histories, I demonstrated that modularity can be restored to a system via the use of multiple simultaneous manipulations, feedback loops, experimental analogs, controlled studies, and the method of subtraction.

I do not take myself to have offered anything like a full or complete account of these strategies. Nevertheless, I do take myself to have broadened the range of cases in biology that mechanistic explanation can account for.

# Future Directions

Among the many questions my dissertation leaves open, I see three broad questions of particular interest. First, can my account correctly handle cases of false inference? Late 19th-century genetics focused on finding the mechanism for heredity, and was

fraught with (as many projects are in their earliest stages) much difficulty. Many candidate mechanisms were proposed, and, in hindsight, nearly all were quite off the mark. Yet most of the proposed mechanisms came with experimental evidence backing them. If my account of the Manipulated Mechanism is on solid footing, then it should offer resources for evaluating the experimental inference underwriting these early proposals relied on, for understanding where the proposals had gone astray, and for explaining why the resulting conclusions were nevertheless somehow plausible.

Second, what happens when experimental conditions are less than ideal? Although one focus of my dissertation is in finding the weakest principles that will support inference about mechanisms, even these very weak assumptions might fail to hold. For example, neuroscientists often assume that distinct brain regions perform distinct functions, and that therefore cognitive activity can be localized to particular brain regions. This assumption is necessary for fMRI studies: In them, two images are taken under slightly different tasks, and the images subtracted. If this assumption holds, then the remainder will indicate which brain regions are responsible for the difference in cognitive activity. But, of course, the assumption that there is a one-to-one correspondence between brain regions and cognitive functionality is suspect at best. What kinds of weaker assumptions could we apply to draw useful inferences from fMRI studies? Thus, one very large research task is to flesh out a set of minimum assumptions for experimental inference, and to test those assumptions against the very messy and complex contexts in which actual experimentation (such as fMRI studies) is performed.

Finally, when constructing models, or deciding how to construct an experiment, are there principles that guide researchers' choice of variables? Theoretical assumptions obviously constrain which variables a researcher will consider, but are there more fundamental assumptions that transcend theory that also place constraints? In

particular, Woodward's modularity condition can be used to distinguish functionally identical, but structurally different models given a set of experimental data. It appears to me that this kind of decision process could be applied to questions of variable selection as well.

# Appendix A

# Non-Linearity and Dynamiticity

What is a non-linear system? A non-linear system is a mathematical construct—a model—comprising a system of non-linear equations.[1]

What, then, is a non-linear equation? A non-linear equation is an equation that is not linear. A linear equation is an equation whose LHS is an additive combination of the paramters on the RHS:

$$f(x_1, \ \ldots \ , x_n) = a_1 x_1 + \ldots + a_n x_n$$

where the $a_i$ are constants. The function $f$ is called a *linear function* if it can be written out in this form. Linear functions exhibit two important properties, *additivity* and *homogeneity*.

**Additivity** $f(x + y) = f(x) + f(y)$.

**Homogeneity** $f(\alpha x) = \alpha f(x)$.

[1]The word 'system' can mean many things, and gets tossed around in this context as though it did not. First, 'system' may be used as shorthand for 'system of equations', which is nothing more than a set of equations taken together. Second, 'system' may be used to refer to a collection of coöperating components in the world, *i.e.* a mechanism. I will use 'system' for the mathematic construct, and 'mechanism' for the thing in the world the system represents.

The function given as an example above satisfies additivity because:

$$\begin{aligned} f(x_1 + y_1, \ \ldots \ , x_n + y_n) &= a_1(x_1 + y_1) + \ldots + a_n(x_n + y_n) \\ &= a_1x_1 + a_1y_1 + \ldots + a_nx_n + a_ny_n \\ &= a_1x_1 + \ldots + a_nx_n + a_1y_1 + \ldots + a_ny_n \\ &= f(x_1, \ \ldots \ , x_n) + f(y_1, \ \ldots \ , y_n) \end{aligned}$$

Likewise does the function satisfy homegeneity:

$$\begin{aligned} f(\alpha x_1, \ \ldots \ , \alpha x_n) &= a_1\alpha x_1 + \ldots + a_n\alpha x_n \\ &= \alpha \left(a_1x_1 + \ldots + a_nx_n\right) \\ &= \alpha f(x_1, \ \ldots \ , x_n) \end{aligned}$$

When plotted to Cartesian coordinates, linear equations will plot as hyperplanes (*i.e.*, in two variables as a line, in three as a plane, *etc.*).

A system of linear equations can be represented as the product of two matrices. If we have $m$ equations in $n$ variables, then one $n \times m$ matrix represents the coffecients in each linear equation, and the other matrix is a column matrix containing the variables:

$$f(\mathbf{X}) = \mathbf{FX} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

A non-linear equation, then, is one that violates either additivity, homogenity, or both. For example, consider the logistic map:

$$x_{n+1} = f(x_n) = rx_n(1 - x_n)$$

The logistic map violates homogeneity:

$$\begin{aligned} f(\alpha x_n) &= r\alpha x_n \left(1 - (\alpha x_n)\right) \\ &= \alpha \left(r x_n (1 - \alpha x_n)\right) \\ &\neq \alpha \left(r x_n (1 - x_n)\right) \end{aligned}$$

and additivity:

$$\begin{aligned} f(x_n + y_n) &= r \left(x_n + y_n\right) \left(1 - (x_n + y_n)\right) \\ &= r \left(x_n + y_n\right) \left(1 - (x_n + y_n)\right) \\ &= \left(r x_n + r y_n\right) \left(1 - (x_n + y_n)\right) \\ &= r x_n (1 - (x_n + y_n)) + r y_n (1 - (x_n + y_n)) \\ &\neq r x_n (1 - x_n) + r y_n (1 - y_n) \end{aligned}$$

What is a dynamical system? A dynamical system is a system of equations, each of which is a dervative of one or more variables with respect to time. That is, it is a system of equations that describes how a set of variables change over time, how they *move*. To take an example from Newtonian mechanics, if the position of a body $x$ is represented by the function $f$ at time $t$, its velocity $v$ is first time-derivative of $x$, and its acceleration $a$ is the first time-derivative of velocity, or the second time-derivative of position:

$$\begin{aligned} x &= f(t) \\ v &= \frac{dx}{dt} = f'(t) \\ a &= \frac{dv}{dt} = \frac{d^2x}{dt^2} = f''(t) \end{aligned}$$

Differential equations can be linear or non-linear. For example, if in the above example $x = -16t^2 + 16t + 32$ (a non-linear function), the velocity is given by $\frac{dx}{dt} = -32t + 16$ (a linear function).

Putting the two together, a non-linear dynamical system is a system of non-linear differential equations.

# Bibliography

Anscombe, G.E.M. (1971). *Causality and Determination*. Cambridge University Press, Cambridge, UK.

Aspect, A., Grangier, P., & Roger, G. (1981). 'Experimental Tests of Realistic Local Theories via Bell's Theorem'. *Physical Review Letters*, 47, pp. 460–463.

Bechtel, W. & Abrahamsen, A. (2005). 'Explanation: A Mechanist Alternative'. *Studies in History and Philosophy of Biology & Biomedical Science*, 36, pp. 421–441.

Bechtel, W. & Richardson, R.C. (1993). *Discovering Complexity: Decomposition and Localization as Scientific Research Strategies*. Princeton University Press, Princeton, NJ.

Beldade, P. & Brakefield, P.M. (2002). 'The genetics and evo-devo of butterfly wing patterns'. *Nature Reviews Genetics*, 3, pp. 442–452.

Bell, J.S. (1964). 'On the Einstein Podolsky Rosen Paradox'. *Physics*, 1, pp. 195–200.

Bogen, J. (2004). 'Analysing Causality: The Opposite of Counterfactual is Factual'. *International Studies in the Philosophy of Science*, 18, pp. 3–26.

Bogen, J. (2005). 'Regularities and Causality; Generalizations and Causal Explanations'. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, pp. 397–420.

Bogen, J. (2008). 'Causally Productive Activities'. *Studies in History and Philosophy of Science*, 39, pp. 112–123.

Bromberger, S. (1966). 'Why-Questions'. In R.G. Colodny, ed., 'Mind and Cosmos: Essays in Contemporary Science and Philosophy', pp. 86–111. University of Pittsburgh Press.

Byrne, M.D., Anderson, J.R., Qin, Y., Bothell, D., Douglass, S.A., & Lebiere, C. (2004). 'An Integrated Theory of the Mind'. *Psychological Review*, 11, pp. 1036–1060.

Cartwright, N. (1999a). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press, Cambridge, UK.

Cartwright, N. (1999b). 'Causal Diversity and the Markov Condition'. *Synthese*, 121, pp. 3–27.

Cartwright, N. (2000). 'Measuring Causes: Invariance, Modularity and the Causal Markov Condition'. Measurement in Physics and Economics 10/00, Centre for Philosophy of Natural and Social Science, London.

Cartwright, N. (2001). 'Modularity: It Can—and Generally Does—Fail'. In M.C. Galavotti, P. Suppes, & D. Costantini, eds., 'Stochastic Causality', pp. 65–84. CSLI Publications.

Cartwright, N. (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*, 53, pp. 411–453.

Cartwright, N. (2004). 'Causation: One Word, Many Things'. *Philosophy of Science*, 71, pp. 805–819.

Chemero, A. & Silberstein, M. (2008). 'After the Philosophy of Mind: Replacing Scholasticism with Science'. *Philosophy of Science*, 75, pp. 1–27.

Cole, K.S. (1949). 'Dynamic electrical characteristics of the squid axon membrane'. *Archives des Sciences Physiologiques*, 3, pp. 253–258.

Cole, K.S. & Moore, J.W. (1960). 'Ionic Current Measurements in the Squid Giant Axon Membrane'. *Journal of General Physiology*, 44, pp. 123–167.

Cooley, T.F. & Leroy, S.F. (1985). 'Atheoretical Macroeconometrics: A Critique'. *Journal of Monetary Economics*, 16, pp. 283–308.

Craver, C.F. (2004). 'Dissociable Realization and Kind Splitting'. *Philosophy of Science*, 71, pp. pp. 960–971.

Craver, C.F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Oxford.

Craver, C.F. (2008). 'Physical Law and Mechanistic Explanation in the Hodgkin and Huxley Model of the Action Potential.' *Philosophy of Science*, 75, pp. 1022–1033.

Craver, C.F. & Bechtel, W. (2007). 'Top-Down Causation Without Top-Down Causes'. *Biology and Philosophy*, 22, pp. 547–563.

Craver, C.F. & Darden, L. (2001). 'Discovering Mechanism in Neurobiology: The Case of Spatial Memory'. In P.K. Machamer, R. Grush, & P. McLaughlin, eds., 'Theory and Method in the Neurosciences', pp. 112–137. University of Pittsburgh Press, Pittsburgh.

Crick, F. (1970). 'Central Dogma of Molecular Biology'. *Nature*, 227, pp. 561–563.

Darden, L. (2001). 'Discovering Mechanisms: A Computational Philosophy of Science Perspective'. In K.P. Jantke & A. Shinohara, eds., 'Discovery Science', pp. 3–15. Springer, New York.

Darden, L. (2002). 'Strategies for Discovering Mechanisms: Schema Instantiation, Modular Subassembly, Forward/ Backward Chaining'. *Philosophy of Science*, 69, pp. S354–S365.

Darden, L. (2006). *Reasoning in Biological Discoveries*. Cambridge University Press, Cambridge, UK.

Darden, L. & Craver, C.F. (2006). 'Strategies in the Interfield Discovery of the Mechanism of Protein Synthesis'. In Darden (2006), pp. 65–97.

Dowe, P. (2000). *Physical Causation*. Cambridge University Press.

Eberhardt, F. & Scheines, R. (2007). 'Interventions and Causal Inference'. *Philosophy of Science*, 74, pp. 981–995.

Eells, E. & Sober, E. (1983). 'Probabilistic Causality and the Question of Transitivity'. *Philosophy of Science*, 50, pp. 35–57.

Fodor, J.A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, Cambridge, MA.

van Gelder, T. (1995). 'What Might Cognition Be, If Not Computation?' *The Journal of Philosophy*, 92, pp. 345–381.

Gibbard, A. & Harper, W. (1981). 'Counterfactuals and two kinds of expected utility'. In W.L. Harper, R.C. Stalnaker, & G. Pearce, eds., 'Ifs: Conditionals, Beliefs, Decision, Chance, and Time', pp. 153–190. Kluwer, Boston.

Gibson, J.J. (1977). 'The Theory of Affordances'. In R. Shaw & J. Bransford, eds., 'Perceiving, Acting, and Knowing', Lawrence Erlbaum Associates, Mahwah, New Jersey.

Glennan, S.S. (1992). *Mechanisms, Models, and Causation*. Ph.D. thesis, Department of Philosophy, University of Chicago, Chicago.

Glennan, S.S. (1996). 'Mechanisms and the nature of causation'. *Erkenntnis*, 44, pp. 49–71.

Glennan, S.S. (1997a). 'Probable Causes and the Distinction between Subjective and Objective Chance'. *Noûs*, 31, pp. 496–519.

Glennan, S.S. (1997b). 'Capacities, Universality, and Singularity'. *Philosophy of Science*, 64, pp. 605–626.

Glennan, S.S. (2002). 'Rethinking Mechanistic Explanation'. *Philosophy of Science*, 69, pp. S342–S353.

Glennan, S.S. (2005). 'Modeling Mechanisms'. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, pp. 443–464.

Glennan, S.S. (2011). 'Singular and General Causal Relations: A Mechanist Perspective'. In P.M. Illari, R. Russo, & J. Williamson, eds., 'Causality in the Sciences', Oxford University Press, Oxford.

Glymour, C. (2002). 'A Semantics and Methodology for *Ceteris Paribus Hypotheses*'. *Erkenntnis*, 57, pp. 395–405.

Glymour, C. (2004). 'James Woodward: Making Things Happen'. *British Journal for the Philosophy of Science*, 55, pp. 779–790.

Goodman, N. (1947). 'The Problem of Counterfactual Conditionals'. *The Journal of Philosophy*, 44, pp. 113–128.

Gopnik, A. & Sobel, D.M. (2000). 'Detecting Blickets: How Young Children Use Information about Novel Causal Powers in Categorization and Induction'. *Child Development*, 71, pp. 1205–1222.

Gopnik, A., Sobel, D.M., Schulz, L.E., & Glymour, C. (2001). 'Causal Learning Mechanisms in Very Young Children: Two-, Three-, and Four-Year-Olds Infer Causal Relations From Patterns of Variation and Covariation'. *Developmental Psychology*, 37, pp. 620–629.

Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., & Danks, D. (2004). 'A theory of causal learning in children: Causal maps and Bayes nets'. *Psychological Review*, 111.

Griffiths, P.E. (2001). 'Genetic Information: A Metaphor in Search of a Theory'. *Philosophy of Science*, 68, pp. 394–412.

Hagmayer, Y. & Sloman, S. (2005). 'Causal Models of Decision Making: Choice as Intervention'. In B.G. Bara, L. Barsalou, & M. Bucciarelli, eds., 'XXVII Annual

Conference of the Cognitive Science Society', Cognitive Science Socirty, Stresa, Italy.

Hagmayer, Y., Sloman, S., Lagnado, D., & Waldmann, M. (2007). 'Causal Reasoning Through Intervention'. In A. Gopnik & L.E. Schulz, eds., 'Causal Learning: Psychology, Philosophy, and Computation', pp. 86–100. Oxford University Press, Oxford.

Haig, B. (2003). 'What Is a Spurious Correlation?' *Understanding Statistics*, 2, pp. 125–132.

Hall, Z.W. (1992). *An Introduction to Molecular Neurobiology*. Sinauer Associates, Sunderland, MA.

Hausman, D.M. & Woodward, J. (1999). 'Independence, Invariance and the Causal Markov Condition'. *British Journal for the Philosophy of Science*, 50, pp. 521–583.

Hausman, D.M. & Woodward, J. (2004a). 'Manipulation and the Causal Markov Condition'. *Philosophy of Science*, 71, pp. 846–856.

Hausman, D.M. & Woodward, J. (2004b). 'Modularity and the Causal Markov Condition: A Restatement'. *British Journal for the Philosophy of Science*, 55, pp. 147–161.

Hempel, C.G. & Oppenheim, P. (1948). 'Studies in the Logic of Explanation'. *Philosophy of Science*, 15, pp. 135–175.

Hille, B., Armstrong, C.M., & MacKinnon, R. (1999). 'Ion Channels: From Idea to Reality'. *Nature Medicine*, 5, pp. 1105–1109.

Hitchcock, C. (2010). 'Probabilistic Causation'. In E.N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2010 edition.

Hodgkin, A.L. & Huxley, A.F. (1952a). 'Currents carried by sodium and postassium ions through the membrane of the giant axon of *Loligo*'. *Journal of Physiology*, 116, pp. 449–472.

Hodgkin, A.L. & Huxley, A.F. (1952b). 'The components of membrane conductance in the giant axon of *Loligo*'. *Journal of Physiology*, 116, pp. 473–496.

Hodgkin, A.L. & Huxley, A.F. (1952c). 'The dual effect of membrane potential on sodium conductance in the giant axon of *Loligo*'. *Journal of Physiology*, 116, pp. 497–506.

Hodgkin, A.L. & Huxley, A.F. (1952d). 'A quantitative description of membrane current and its application to conduction and excitation in nerve'. *Journal of Physiology*, 117, pp. 500–544.

Hodgkin, A.L., Huxley, A.F., & Katz, B. (1952). 'Measurement of current-voltage relations in the membrane of the giant axon of *Loligo*'. *Journal of Physiology*, 116, pp. 424–448.

Hume, D. (1777). *An Enquiry Concerning Human Understanding*. Project Gutenberg, Salt Lake City, 10th edition.

Huxley, A.F. (2002). 'From Overshoot to Voltage Clamp'. *Trends in Neuroscience*, 25, pp. 553–558.

Kauffman, S. (1970). 'Articulation of Parts Explanation in Biology and the Rational Search for Them'. In R. Buck & R. Cohen, eds., 'Boston Studies in the Philosophy of Science', volume VIII. D. Reidel, Dordrecht.

Koester, J. & Siegelbaum, S.A. (2000). 'Propagated Signalling: The Action Potential'. In E.R. Kandel, J.H. Schwartz, & T.M. Jessell, eds., 'Principles of Neural Science', chapter 9, pp. 150–170. McGraw-Hill, fourth edition.

Kyburg, H. (1965). 'Discussion: Salmon's Paper'. *Philosophy of Science*, 32, pp. 147–151.

Loewi, O. & Navratil, E. (1926). 'Über humorale Übertragbarkeit der Herznervenwirkung'. *Pflügers Archiv European Journal of Physiology*, 214, pp. 678–688.

Lotka, A.J. (1925). *The Elements of Physical Biology*. Williams & Williams Co., Baltimore.

Machamer, P.K. (2004). 'Activities and Causation: The Metaphysics and Epistemology of Mechanisms'. *International Studies in the Philosophy of Science*, 18, pp. 27–39.

Machamer, P.K., Darden, L., & Craver, C.F. (2000). 'Thinking about Mechanisms'. *Philosophy of Science*, 67, pp. 1–25.

Menzies, P. & Price, H. (1993). 'Causation as a Secondary Quality'. *The British Journal for the Philosophy of Science*, 44, pp. 187–203.

Mill, J.S. (1843). *A System of Logic, Ratiocinative and Inductive*. John W. Parker.

Miller, R., Barnet, R., & Grahame, N. (1995). 'Assessment of the Rescorla-Wagner model'. *Psychological bulletin*, 117, pp. 363–386.

Montaña, J. (2009). 'Break the Toaster: Manipulationist Descriptions of Mechanisms'. The International Society for the History, Philosophy, and Social Studies of Biology, Brisbane, Australia, July, 2009.

Norton, J.D. (2003). 'Causation as Folk Science'. *Philosophers' Imprint*, 3, pp. 1–22.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York.

Psillos, S. (2004). 'A Glimpse of the Secret Connexion: Harmonizing Mechanisms with Counterfactuals'. *Perspectives on Science*, 12, pp. 288–319.

Rescorla, R. & Wagner, A. (1972). 'A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement'. In A. Black & W. Prokasy, eds., 'Classical Conditioning II', pp. 64–99. Appleton-Century-Crofts.

Salmon, W.C. (1979). 'Review: Propensities: A Discussion Review'. *Erkenntnis*, 14, pp. 183–216.

Salmon, W.C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton.

Salmon, W.C. (1989). *Four Decades of Scientific Explanation*. University of Minnesota Press, Minneapolis.

Salmon, W.C. (1994). 'Causality without Counterfactuals'. *Philosophy of Science*, 61, pp. 297–312.

Salmon, W.C. (1997). 'Causality and Explanation: A Reply to Two Critiques'. *Philosophy of Science*, 64, pp. 461–477.

Salmon, W.C. (1998). *Causality and Explanation*. Oxford University Press, Oxford.

Sands, Z., Grottesi, A., & Sansom, M.S. (2005). 'Voltage-Gated Ion Channels'. *Current Biology*, 15, pp. R44 – R47.

Simon, H.A. (1962). 'The Architecture of Complexity'. *Proceedings of the American Philosophical Society*, 106, pp. 467–482.

Sloman, S.A. & Lagnado, D.A. (2005). 'Do We "do"?' *Cognitive Science*, 29, pp. 5–39.

Sober, E. (2001). 'Venetian sea levels, British bread prices, and the principle of the common cause'. *The British Journal for the Philosophy of Science*, 52, pp. 331–346.

Sober, E. & Lewontin, R.C. (1982). 'Artifact, Cause and Genic Selection'. *Philosophy of Science*, 49, pp. 157–180.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction and Search*. Springer Lecture Notes in Statistics. MIT Press, Harvard, MA.

Steel, D. (2006). 'Comment on Hausman and Woodward on the Causal Markov Condition'. *British Journal for the Philosophy of Science*, 57.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland, Amsterdam.

Suppes, P. (1986). 'Non-Markovian Causality in the Social Sciences with some Theorems on Transitivity'. *Synthese*, 68, pp. 129–140.

Tabery, J. (2009). 'Difference Mechanisms: Explaining Variation with Mechanisms'. *Biology and Philosophy*, 24, pp. 645–664. 10.1007/s10539-009-9161-2.

Tabery, J.G. (2004). 'Synthesizing Activities and Interactions in the Concept of a Mechanism'. *Philosophy of Science*, 71, pp. 1–15.

Thorne, R.E. & Thomas, G.L. (2008). 'Herring and the "Exxon Valdez" oil spill: An investigation into historical data conflicts'. *International Council for the Exploration of the Sea Journal of Marine Science*, 65, pp. 44–50.

Volterra, V. (1926). 'Fluctuations in the abundance of a species considered mathematically'. *Nature*, 118, pp. 558–560.

Wagner, G.P., Pavlicev, M., & Cheverud, J.M. (2007). 'The road to modularity'. *Nature Reviews Genetics*, 8, pp. 921–931.

Wikipedia (2010). 'HIV — Wikipedia, The Free Encyclopedia'. [Online; accessed 26-February-2010].

Wolf, S. & Hardy, J.D. (1941). 'Studies on pain. Observations on pain due to local cooling and on factors involved in the "cold pressor" effect'. *Journal of Clinical Investigation*, 20, pp. 521–533.

Woodward, J. (1997). 'Explanation, Invariance, and Intervention'. *Philosophy of Science*, 64, pp. S26–S41.

Woodward, J. (1999). 'Causal Interpretation in Systems of Equations'. *Synthese*, 121, pp. 199–257.

Woodward, J. (2000). 'Explanation and Invariance in the Special Sciences'. *British Journal for the Philosophy of Science*, 51, pp. 197–254.

Woodward, J. (2002). 'What is a Mechanism? A Counterfactual Account'. *Philosophy of Science*, 69, pp. S366–S377.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.

Woodward, J. & Hitchcock, C. (2003a). 'Explanatory Generalizations, Part I: A Counterfactual Account'. *Noûs*, 37, pp. 1–24.

Woodward, J. & Hitchcock, C. (2003b). 'Explanatory Generalizations, Part II: Plumbing Explanatory Depth'. *Noûs*, 37, pp. 181–199.

von Wright, G.H. (1971). *Explanation and Understanding*. Cornell University Press, Ithica, NY.

Yuan, C.S., Karrison, T., Wu, J.A., Lowell, T.K., Lynch, J.P., & Foss, J.F. (1998). 'Dose-related effects of oral acetaminophen on cold-induced pain: A double-blind, randomized, placebo-controlled trial'. *Clinical Pharmacology & Therapeutics*, 63, pp. 379–383.