6-2022

Actual Causation: Apt Causal Models and Causal Relativism

Jennifer R. McDonald
The Graduate Center, City University of New York

ACTUAL CAUSATION:
APT CAUSAL MODELS AND CAUSAL RELATIVISM

by

JENNIFER MCDONALD

A dissertation submitted to the Graduate Faculty in Philosophy in partial fulfillment of the requirements for the degree of the Doctor of Philosophy, The City University of New York

2022

ACTUAL CAUSATION:
APT CAUSAL MODELS AND CAUSAL RELATIVISM

by

Jennifer McDonald

This manuscript has been read and accepted for the Graduate Faculty in Philosophy
In satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date	David Papineau Chair of Examining Committee
Date	Nickolas Pappas Executive Officer

Supervisory Committee:

Stephen Neale
Graham Priest
Jonathan Schaffer
Michael Strevens

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

ACTUAL CAUSATION: APT CAUSAL MODELS AND CAUSAL RELATIVISM

by Jennifer McDonald

Advisor: David Papineau

This dissertation begins by addressing the question of when a causal model is *apt* for deciding questions of actual causation with respect to some target situation. I first provide relevant background about causal models, explain what makes them promising as a tool for analyzing actual causation, and motivate the need for a theory of aptness as part of such an analysis *(Chapter 1)*. I then define what it is for a model on a given interpretation to be *accurate of*, that is, say only true things about, some target situation. This involves a systematization of various representational principles mentioned and/or discussed throughout the literature into a method of interpretation, which I propose be taken as standard *(Chapter 2)*. Next, I explain and address two reasons for which accuracy as I've defined it is insufficient for aptness. The first reason – already discussed in the literature – is the problem of structural isomorphs. In response, I propose the aptness condition of Explicit Partial Mediation *(Chapter 3)*. The second reason – which has yet to be noticed – is the problem of the indeterminacy of accuracy. As I demonstrate, a model is accurate of a target situation only relative to a set of background possibilities – what I call a *modal profile*. It follows that a model represents a situation only relative to some modal profile or other. I go on to discuss the ramifications of this observation for a theory of actual causation in terms of models. I argue that the relativity be taken at face value and built into our metaphysical account of causation, resulting in a view that I call causal relativism *(Chapter 4)*. I explore one advantage of this view in detail: that the resulting account can defend the principle of strong proportionality against several objections *(Chapter 5)*. Finally, I apply the earlier discussion of aptness to attempts to provide a semantics of counterfactuals in terms of causal models – an *interventionist semantics*. I show how just as a similarity semantics relies on an opaque notion of *similarity*, an interventionist semantics relies on an analogous notion of *aptness*. The challenge of articulating aptness thus undermines the claim that an interventionist semantics avoids representational problems inherent in a similarity semantics *(Chapter 6)*. I close with a recap and suggestions for future research *(Chapter 7)*.

ACKNOWLEDGMENTS

> “And therefore never send to know for whom the bell tolls; it tolls for thee.”
>
> (Donne, 1624)

No man is an island, and even less so is a graduate student. For my journey thus far through academia and the production of this dissertation in particular, I owe a debt of gratitude to scores of people and to several institutions. First among them is my advisor, David Papineau, without whom the quality of this work, of my philosophical ability, and of my graduate experience would be substantially reduced. David consistently pushed me to improve both my philosophical writing and skill in conversation, and he set an equally high standard for whether the work is clear, accessible, and rigorous as for whether it is interesting and worth our time. I cannot begin to express my thanks for his generosity and attention.

I would also like to express my deepest appreciation to the members of my committee. The exacting and insightful feedback that I received from Jonathan Schaffer and Michael Strevens has been invaluable in the improvement of this document and in my abilities. I am thankful to them, as well, for their kindness and humor. My thinking through these issues and many others was also greatly aided by discussions and courses taken with Stephen Neale and Graham Priest. I am grateful to them for their penetrating feedback on my work and their demonstration of what deep and interesting philosophy looks like.

I am further indebted to Brad Weslake and Samuel Lee for lively conversation, which has been instrumental in shaping and deepening my understanding on these topics. I'd like to acknowledge helpful suggestions I received on written work pertaining to this dissertation from each of Laura Gradowski, Zoey Lavallee, Jessie McCormack, and Fiona Schick. Finally, I wish to express appreciation for the audiences at the following talks: Philosophy of Science Biennial Conference (Seattle, 2018); Society for the Metaphysics of Science Annual Conference (Milan, 2018 & Toronto, 2019); Logic and Metaphysics Workshop (NYC, Spring 2019 & Fall 2020); Philosophy of Language Workshop (NYC, Spring 2020); Philosophy of Logic Seminar (Spring 2021); CUNY Brown Bags (Spring 2018, Spring 2019, Fall 2020, & Spring 2021); and University of Illinois at Urbana-Champagne Philosophy Graduate Conference (Spring 2021). The opportunity to present and to field challenging feedback on each of these occasions contributed greatly to the quality of this work and to my enjoyment in crafting it.

There are many whose intelligence, wit, kindness, and drive have guided and motivated me along the way. A great many thanks to Harjit Bhogal, Shannon Brick, Noëll Carroll, Alexander Franklin, Miranda Fricker, Justin Garson, Jonathan Gilmore, Peter Godfrey-Smith, John Greenwood, Joel David Hamkins, Daniel Harris, Paul Horwich, Jonathan Kwan, Colin Marshall, Josh May, Vincent Peluce, Kate Ritchie, David Rosenthal, Ian Rumfitt, Jules Salomone, Hagop Sarkissian, Steven Schiffer, Alex Silk, Derek Skillings, Maja Spener, Naomi Spector, Scott Sturgeon, Kathryn Tabb, Peter Tan, Thomas Teufel, Iakovos Vasiliou, Alastair Wilson, Kathryn Wojtkiewicz, and Crispin Wright. Thanks also to whomever I have surely yet accidentally left off this list.

A mind does not run on spirit alone. I am honored and grateful to have received financial support throughout this process from the Graduate Center CUNY – via the Graduate Center Fellowship, Dean Harrison Dissertation Fellowship, and McGraw Dissertation Award – and from the Society for Christian Philosophers/ Templeton Foundation – via the Graduate Student Cross-Training Fellowship.

A very special thanks is due Justin Clarke-Doane. Without his guidance and reassurance, the completion of my dissertation would not have been possible. And without his incisive and challenging feedback, the work would be greatly diminished. For keeping me sane, and for his warmth, his humor, and his love, I am grateful beyond words.

Finally, I wish to thank my family and friends for their unwavering support and belief in me. To my mom, in particular, whose strength, success, and grace during her lifetime has been an abiding source of inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Actual Causation and the Promise of Causal Models

> We live in exciting times. By 'we' I mean philosophers studying the nature of causation. The past decade or so has witnessed a flurry of philosophical activity aimed at cracking this nut, and, surprisingly, real progress has been made…. [T]here has been increasing philosophical interest in the techniques of causal modeling developed and employed within fields such as economics, epidemiology, and artificial intelligence.
>
> (Hitchcock, 2001, p. 273)

§1.1 Introduction

Actual causation, also called *token causation* or *singular causation*, is the relation that holds between two particular things when the first causes the second. This relation is reflected in causal claims such as:

> Mount Vesuvius erupting in 79 AD caused the city of Pompei to be buried in ash.
>
> Cory skipping class on Wednesday caused her to miss the test.
>
> The cat knocking the vase off the table caused it to break.

A precise account of actual causation has proven elusive. But recent progress seems to have been made utilizing the framework of structural equation models, or *causal models* for short.

These mathematical models come from the special sciences (econometrics, statistics, computer science, etc.), where they have been developed and refined over several decades to better understand causal structure, discover causes, and make predictions (Pearl, 2000; Spirtes et al., 1993). They represent in a formally sophisticated way the intuitive idea that causation is difference-making. A *structural equation model* is a set of variables, taken to represent the causal relata, and a set of asymmetric functional equations defined over them, taken to represent dependency relations holding between the relata.

In general, an account of actual causation in terms of these models provides necessary and sufficient conditions for an actual causal relation holding between two particular things in terms of properties of a model that appropriately represents those things – an *apt* model. There are two stages to the provision of such an account. The first answers the question of *which properties* of a model are the right ones for identifying with an actual causation relation. Answering this gives a recipe by which one can read actual causation relations off of an apt model. A very simple recipe would be that *c* is an actual cause of *e* just in case intervening on an apt model to change the value of the variable that represents *c* leads to a change in the value of the variable that represents *e*. For example, the eruption of Mount Vesuvius in 79 AD is an actual cause of the city of Pompei being buried in ash in 79 AD just in case intervening on an apt model to change the value of the eruption variable leads to a change in the value of the buried-in-ash variable. Of course, this recipe only works when causal dependence lines up with counterfactual dependence, and does not cover cases of redundant causation. While this question of how to articulate the recipe has received considerable attention, it remains less than entirely settled.[1]

The second stage will be my real focus. This is to provide a principled account of what qualifies a model or class of models as *apt* – eligible to be plugged into the first stage. We want apt models to be those where the application of the correct recipe delivers only true causal verdicts. But which ones are these? This question is difficult and often bracketed. [2] Work on the first stage progresses largely by a reliance on "natural" models. Arguably, though, we find these models natural because they capture what we already understand about the causal structure of the situation in question. Thus, if the models are to independently illuminate the nature of causation, an articulation of aptness is crucial.

The rest of this introductory chapter provides relevant background. §1.2 overviews the formal apparatus of structural equation models and directed acyclic graphs. I will use the expression "causal model" throughout to refer to either or both of these types of models. §1.3 provides a particular recipe of actual causation and demonstrates why SEM definitions seem to be a promising new means with which to define actual causation. Finally, §1.4 explains the need for an account of aptness and clarifies what role aptness must play.

§1.2 The Formal Apparatus of Causal Models

[1] I will say a bit more about this in §1.3. See (Halpern & Pearl, 2005; Hitchcock, 2001; Weslake, 2015; Woodward, 2003) for various causal model definitions of actual causation.

[2] Of course, it has not been universally bracketed. For good work on the issue, see (Blanchard & Schaffer, 2017; J. Halpern & Hitchcock, 2010; J. Y. Halpern, 2016b; Hitchcock, 2001, 2012; Woodward, 2016).

This section provides a formal overview of the kinds of causal models in question – *structural equation models (SEMs)* and corresponding *directed acyclic graphs (DAGs)*. SEMs can be used to represent either deterministic or probabilistic systems, and token or type level structures of either kind. For expository purposes, I will make two simplifying assumptions. First, I will focus only on how these models represent deterministic systems, setting aside probabilistic ones. Second, for the purposes of representing actual causation, I will focus on token level models. There is undoubtedly significant overlap between how models are used to represent probabilistic and deterministic systems, and token- and type-level ones. However, I leave this bridge to be constructed in future work.

The reader familiar with work on causal models may notice that I take care to carve the formalism of these models away from what they can be taken to represent. In other words, I distinguish between the formalism and what might be called the representational semantics.[3] Often, these are blended together in relatively benign ways that facilitate uptake and don't otherwise muddy inquiry into the intended topic. However, since this dissertation aims to clarify precisely what the representational content for these models is – what governs the process of assigning real-world content to a model – it therefore makes sense to separate them. This section lays out the formalism. Chapter 2 will articulate various principles of representation, systematizing them into what I propose be taken as the standard representational semantics, at least for the purposes of defining actual causation.

[3] "Representational" is not redundant here. The formalism of a structural equation model is itself a layer of semantics over the syntax of the system of numerals. Thanks to David Papineau for this point.

1.2.a The Formalism of SEMs

The formalization of causal models with which I'll work follows Halpern (2000) and Blanchard and Schaffer (2017) and is adapted only slightly. A causal model comprises three levels of structure, $\{\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{A}}\}$: the Signature, the Linkage, and the Assignment. The *Signature* is the collection of variables. It includes a set of exogenous variables, a set of endogenous variables, and a function that maps to each variable a range of possible values, where each range has at least two members. Formally, $\boldsymbol{\mathcal{S}} = < \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{R} >$, where $\boldsymbol{U}$ is the set of exogenous variables, $\boldsymbol{V}$ is the set of endogenous variables, and $\boldsymbol{R}$ is a function mapping a set of values with at least two members to each variable $X: X \in (\boldsymbol{U} \cup \boldsymbol{V})$. While the method for how content is assigned to a model will be explored in greater depth in the next chapter, it may be helpful here to mention that, intuitively, variables represent possible causal relata. For example, a binary variable can represent whether a particular event might occur or not. A many-valued variable can represent a range of possible masses of a particular object.

The second level of structure – the *Linkage* – is a set of functional equations, $\boldsymbol{\mathcal{L}}$, defined over this set of variables. The form of each equation is such that a single variable appears on the left-hand side, and some function on some subset of variables excluding the left-hand one appears on the right. The form of the right-hand function can be whatever suits the needs of the modeler. We might have, for example, '$Y := 3/2X + Z$', '$R := S \vee T$', or '$Z := max(R, X)$'. The functional equations of a model are asymmetric. They stipulate what value the left-hand variable – called the *child variable* – will take for any combination of values of the right-hand variables – called the *parent variables*, when these variables are set to their values by *intervention* – a technical term I will define shortly. For example, consider the sample model in *Figure 1*.

$\mathcal{S}$ = $\boldsymbol{U} = \{X\}$
$\boldsymbol{V} = \{Y, Z\}$
$\boldsymbol{R} = f(X_i) = \{1, 0\}$

$\mathcal{A}$ = (EQ1) $X = 1$

$\mathcal{L}$ = (EQ2) $Y := (1 - X)$
(EQ3) $Z := \max(X, Y)$

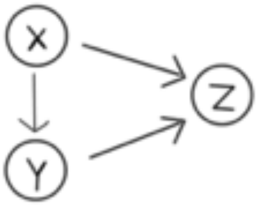


Figure 1. $\boldsymbol{\mathcal{M}_1}$

EQ3 of $\mathcal{M}_1$, for example, says that the value of *Z* is determined by the values of *X* and *Y* in that the value of *Z* will be the maximum value of either *X* or *Y*. (A quick bit of terminology: this is the *Z-equation*. For reasons laid out shortly in §1.2.c, any model I will be considering has at most one such equation for any given variable.) The formal nature of the equations permits the model to state precisely what happens to certain variables as the values of other variables change. They therefore capture quite naturally information about how the target system would evolve under various alterations. Intuitively, these equations represent dependency relations between the variables. EQ3 specifies that a dependency relation holds between what is represented by the variables *X* and *Y* and what is represented by Z, defined by the maximum function.

Finally, the third level of structure, the *Assignment*, sets each exogenous variable to some particular value from its range of possible values.[4] Since the Assignment is standardly taken to represent the initial conditions of the target system, it is also sometimes called the *Context*. Formally, it is a function, $\boldsymbol{\mathcal{A}}$, that, to every variable U_x: $U_x \in \boldsymbol{U}$, maps a value u_x: $u_x \in \boldsymbol{R}(U_x)$. Each such mapping can be represented as a constant equation, and I will take these constant equations to be part of the model's complete set of structural equations, alongside the functional equations from the Linkage. This slight adaptation of the formalism permits interventions on exogenous variables, for reasons I will now explain.

The notion of intervention is a technical one and, as I use the term here, a merely formal one.[5] An *intervention* on a variable, *X*, is a targeted operation that forces *X* and only *X* to one of its possible values, breaking *X*'s dependence on its parent variables, if it has any, in the process, and otherwise leaving the model as is. I follow Pearl (2000) in treating an intervention on a model as an operation that produces a sub-model.[6] More precisely, an intervention, $I_{X=xi}$, on a variable, *X*, in a model, $\mathcal{M}_i$ = {$\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{A}}$}, produces a sub-model, $\mathcal{M}_{i,X=x_i}$ = {$\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{L}}', \boldsymbol{\mathcal{A}}'$} in which {$\boldsymbol{\mathcal{L}}'$, $\boldsymbol{\mathcal{A}}'$} is the same as {$\boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{A}}$} except that the *X*-equation is replaced by 'X=x_i'. Such an operation renders *X* independent of its parent variables, if it has any, but preserves the

[4] There is some variation in the literature as to whether the Assignment is considered a constituent of a model or not. Were we employing these models to represent type-level causation, it would be expedient to individuate models solely on the basis of the other two constituents, $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{L}}$. For actual causation, where the focus is on concrete situations, it makes more sense to build $\boldsymbol{\mathcal{A}}$ into the model. While this choice shapes the form of one's theory, it has no philosophical upshot.

[5] A richer notion of intervention is presented in (Woodward, 2003), which will be relevant only later on when I discuss the representational semantics of these models.

[6] See also (Briggs, 2012).

dependency structure down stream of X. For example, the intervention $I_{Y=0}$ on $\mathcal{M}_1$ would produce the sub-model in *Figure 2*.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{X\}$
$\boldsymbol{V} = \{Y, Z\}$
$\boldsymbol{R} = f(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $Y = 0$
(EQ3) $Z := \max(X, Y)$

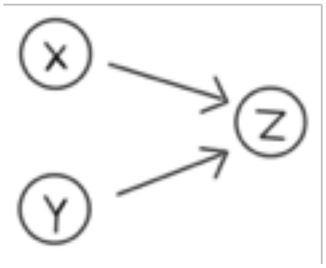


Figure 2. $\boldsymbol{\mathcal{M}_{1,Y=0}}$

1.2.b Directed Graphs

It can be helpful to represent structural equation models with directed graphs. Such graphs abstract away from the precise nature of the represented dependencies, indicating only the simple fact that the dependencies exist. A directed graph is a set of nodes, or vertices, and a set of directed edges (drawn as arrows) between those nodes. The nodes can be taken to correspond to the variables of a SEM and the directed edges can be taken to correspond to the equations of a SEM, drawn from parent variables to child variables (see *Figure 1*, above).

1.2.c Recursiveness and Acyclicity

A final point of note on the formalism. For defining causation, it's standard to focus on *recursive* SEMs. In this context, recursive means that the equations can be ordered such that once a variable appears on the right-hand side it does not again appear on the left-hand side. This effectively rules out cycles. So, recursive SEMs correspond to directed *acyclic* graphs (DAGs). It follows that while exogenous variables get their values from the Assignment, endogenous variables are assigned a unique value as a result of the Assignment and the Linkage.

§1.3 A Recipe for Actual Causation

The SEM framework is then employed to provide a recipe for reading relations of actual causation off of a model. There are various such recipes in the literature.[7] But since the focus of this dissertation is on articulating aptness rather than on refining the recipe, I will adopt a relatively simple recipe for my purposes, which is originally due to Hitchcock (2001, p. 290).

In order to introduce this, I will first need to quickly define the notion of a *directed path* in a model. A *directed path* in a SEM, hereafter a *path*, is a set of ordered variables, <$X_1, X_2, X_3, \ldots, X_i$>, such that X_1 is a parent of X_2, X_2 is a parent of X_3, …, and X_{i-1} is a parent of X_i. The set of nodes corresponding to such variables in a corresponding DAG are such that the arrows between them all point in the same direction. With this in place, I can now lay out the following recipe:

[7] See (Gallow, forthcoming; Halpern & Pearl, 2005; Hitchcock, 2001; Weslake, 2015; Woodward, 2003) for various causal model definitions of actual causation.

($\boldsymbol{AC - relative}$) $X = x$ is an actual cause of $Y = y$ in $\mathcal{M}_i$ just in case…

AC1) $X = x$ and $Y = y$ in $\mathcal{M}_i$.

AC2) There is a directed path P_i in $\mathcal{M}_i$ from X to Y and an assignment of values to the set of variables off P_i such that the following are true:

(a) Were the off-path variables (call these $\vec{Z}$) set by intervention to the specified assignment (call this $\vec{z}$), then the variables on P_i would still have taken their actual values.

(b) Were $\vec{Z} = \vec{z}$ and $X = x$ set by intervention, then $Y = y$.

(c) Were $\vec{Z} = \vec{z}$ and $X = x_i$ set by intervention, where $x_i \neq x$, then $Y = y_i$, where $y_i \neq y$.

$\boldsymbol{AC}$1 is an actuality condition. It requires that the model represent $X = x$ and $Y = y$ as actually occurring. Combined with the model's aptness, which requires that the model say only true things, this ensures that the cause and effect do actually occur.

$\boldsymbol{AC}$2 is the causal condition. $\boldsymbol{AC}$2 says that there must be a path between the putative cause and effect such that when all off-path variables are held fixed at values that satisfy $\boldsymbol{AC}$2a, then intervening to set the putative cause as occurring will result in the effect occurring ($\boldsymbol{AC}$2b), and intervening to set some alternative to the putative cause will result in some alternative to the effect occurring ($\boldsymbol{AC}$2c). The condition of $\boldsymbol{AC}$2a requires of the setting of off-path variables that it preserve the actual values of the on-path variables.

$\boldsymbol{AC-relative}$ is just one version among many. Different iterations of the recipe have been devised in response to different problem cases such as the infamous cases of *redundant causation* – cases where the effect doesn't straightforwardly counterfactually depend on the cause, which I'll discuss in the next section. The different versions principally disagree on what constitutes a *permissible* setting of off-path variables. $\boldsymbol{AC-relative}$ requires only that the setting of off-path variables satisfy $\boldsymbol{AC}$2a, but there is reason to think this won't ultimately work (Weslake, 2015). Despite this, I will employ the fairly simplistic $\boldsymbol{AC-relative}$ throughout since refining the recipe is not my focus, and the ways in which this particular recipe fails to capture actual causation will not affect any of my arguments.

§1.4 The Promise of Causal Models

The ability of a causal model to distinguish between on-path and off-path variables, holding them fixed in different ways, is precisely what makes causal models promising. This is what allows causal models to distinguish between two distinct paths of influence between the same two things. By distinguishing, the causal model framework has the vocabulary to analyze the activity along each path separately from activity along the other. This greatly advantages the structural equation framework over a simple counterfactual framework. Hitchcock explains,

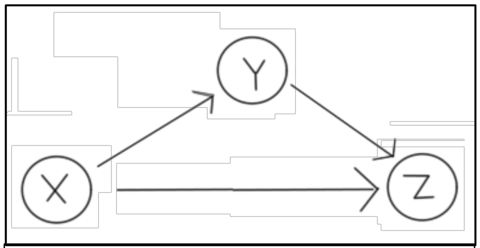


Figure 3.

> An arrow from *X* to Z [as, for example, in *Figure 3*] thus means that the value of *Z* can depend counterfactually upon the value of *X*, *even holding fixed the value of Y*. The natural causal interpretation of this counterfactual is that the value of *X* can have an effect on the value of *Z* over and above the effect it has in virtue of causing the value of *Y*. There are two routes whereby *X* influences *Z*; one which runs through *Y*, and one direct route which bypasses *Y*. The overall effect of *X* on *Z* will depend upon both of these routes. (2001, p. 285)

Hitchcock goes on to exploit this distinction between on-path and off-path variables to address issues surrounding the transitivity of causation (Hitchcock, 2001). Exploiting this distinction is also how SEM definitions of actual causation handle the infamous problem of redundant causation. Redundant causation occurs when there is actual causation without counterfactual dependence. Take, as an example, the following paradigmatic case of what is called early preemption:

Early Preemption Suzy and Billy are throwing rocks at a window. Suzy throws a rock at the window, the rock hits the window, and the window shatters. Billy refrains from

> throwing when he sees Suzy throw. But had Suzy not thrown, then Billy would have. And had Billy thrown, the window would still have shattered.

Here, Suzy's throw causes the window to shatter, despite the fact that had Suzy not thrown, then Billy would have, and the window still would have shattered. Thus, the window's shattering does not counterfactually depend on Suzy's throw in a straightforward way. SEM definitions of actual causation respond to this by pointing out that counterfactual dependence between the effect and the redundant cause can be revealed by conditioning on certain other features of the situation – namely, *off-path* features. Hold fixed the fact that Billy doesn't throw, for example, and we recover the dependence of the window's shattering on Suzy's throw. Had Suzy not thrown *and had Billy not thrown*, then the window would not have shattered.

To see this in action, we can use our sample model, $\mathcal{M}_1$, to represent **Early Preemption**. Let's assume the following natural interpretation on $\mathcal{M}_1$:

$$\mathcal{I}(\mathcal{M}_1)_{EP} : \qquad X = \begin{cases} 1\ if\ Suzy\ throws\ a\ rock \\ 0\ if\ Suzy\ doesn't\ throw \end{cases}$$

$$Y = \begin{cases} 1\ if\ Billy\ throws\ a\ rock \\ 0\ if\ Billy\ doesn't\ throw \end{cases}$$

$$Z = \begin{cases} 1\ if\ the\ window\ shatters \\ 0\ if\ the\ window\ doesn't\ shatter \end{cases}$$

According to ***AC* – *relative***, *X = 1* is an actual cause of *Z = 1* in $\mathcal{M}_1$. ***AC1*** is satisfied, since *X = 1* and *Z = 1*. ***AC2*** is satisfied, with the relevant path being {*X*, *Z*} and the relevant setting of off-path variables being *Y = 0*. First, ***AC2****a* is satisfied, since when *Y = 0*, the variables on the path keep their actual values of *X = 1* and *Z = 1*. Next, ***AC2****b* is satisfied, since when *Y = 0* and *X = 1* then *Z = 1*. Finally, ***AC2****c* is satisfied. When *Y = 0* and *X = 0*, then *Z = 0*.

On $\mathcal{I}(\mathcal{M}_1)_{EP}$, this means that Suzy's throw (*X = 1*) is an actual cause of the window shattering (*Z = 1*) and, crucially, Billy's throw is not considered an actual cause of the window shattering. This for the simple reason that it's not the case that *Y = 1* in $\mathcal{M}_1$, and so ***AC1*** goes unsatisfied.

§1.5 Aptness

Notice that ***AC* – *relative*** defines actual causation only relative to a model. In order to transform recipes like ***AC* – *relative*** into a general definition of causation, common practice is to quantify over a set of apt models. The standard way to go seems to be to existentially quantify (Blanchard & Schaffer, 2017; Hitchcock, 2001; Weslake, 2015), which produces the following SEM definition of actual causation:

> $(\boldsymbol{AC - simpliciter})^{-}$ *c* is an actual cause of *e* just in case there is an apt model, $\mathcal{M}_i$, which represents *c* as *X* = *x* and *e* as *Y* = *y*, and delivers the ***AC* – *relative*** verdict that *X* = *x* is an actual cause of *Y* = *y*.

Of course, this quantifier selection is not logically required. One could instead universally quantify over apt models (N. Hall, 2007).[8] Alternatively, one could manage by only allowing one model to be apt for any given target situation.[9] But this isn't a substantive issue. Regardless of which of these options is chosen, more will be required in order to provide a philosophically satisfying account of actual causation. In particular, any theory will need to say something about aptness, and one's choice between the above options will simply affect what needs to be said. Allow me to explain.

Which quantifier one chooses will determine what requirements must be placed on aptness. An existential quantifier will call for aptness requirements that rule out models which mistakenly witness actual causation, where there isn't any causation. A universal quantifier will call for aptness requirements that rule out models which mistakenly witness the absence of actual causation, where there is causation. Alternatively, a definition could eschew quantification and only allow one model per target situation. But then we would need a different notion of aptness to deliver a unique model.

So, everyone has an aptness problem. A natural starting question here is: why can't we just quantify over *literally* all models? The immediate response is that models will at least need to be *accurate* – that is, a model will need to get the target situation roughly right. The first step in articulating aptness, then, is to define accuracy. The next chapter tackles this project.

[8] Indeed, one could quantify in any number of ways. Thanks to Jonathan Schaffer for drawing my attention to this point.

[9] Note that nowhere in the literature is a unique-apt-model view defended. But it's a natural option to adopt for those who provide a strictly model-relative account of causation– such as (J. Y. Halpern, 2016a; J. Y. Halpern & Hitchcock, 2015; Halpern & Pearl, 2005).

Before I begin, however, it will help to briefly clarify the nature of the property of aptness. Aptness is not simply a property of models. A causal model is a *mathematical object*. As such, it has no real-world content until given an interpretation. It therefore doesn't make sense to say that a model is apt or inapt on its own, but only *on an interpretation*. Yet this is still too simple. A model is apt on an interpretation only *relative to a target situation*. So, aptness is really a relation between three things – a model, an interpretation, and a situation. The complexity of aptness is often overlooked in the literature, but it will serve to keep this clear. This calls for a slight update to the earlier SEM definition of actual causation:

> ($\boldsymbol{AC - simpliciter}$) c is an actual cause of e just in case there is an apt model-interpretation pair, <$\mathcal{M}_i$, $\mathcal{I}(\mathcal{M}_i)$>, where $\mathcal{I}(\mathcal{M}_i)$ represents c as $X = x$ and e as $Y = y$, and $\mathcal{M}_i$ delivers the ***AC – relative*** verdict that $X = x$ is an actual cause of $Y = y$.

CHAPTER 2

A Guide to Interpreting Causal Models for Actual Causation

> It is an excellent question, inadequately addressed in the literature, precisely what principles should guide the construction of a causal model.
>
> (Paul & Hall, 2013, pp. 18–19)

> [R]elatively little has been done to get clear about what exactly someone commits themselves to when they endorse one of these models – what exactly, that is, a structural equations model *says* about the world.
>
> (Gallow, 2016, p. 160)

§2.0 Abstract This chapter begins by systematizing and explicating the principles governing what a causal model says, as indexed to the purpose of representing actual causation. It offers a thorough articulation of the process of model interpretation. It then defines what it means for a model to be *accurate* of a situation on a given interpretation. I propose that this method be taken as standard.

§2.1 Introduction

In defining what it is for a causal model to be *accurate*, the question first arises as to how it relates to the world at all. Thus, in order to articulate aptness we need to first be clear on what a model says. The answer to this of course depends on our method of interpretation. So, an account of aptness cannot get off the ground without first settling on a method of interpretation for causal models. Yet there is no explicit and comprehensive method of interpretation in the literature. This chapter makes some progress in this direction. While there does not seem to be an official standard, I aim in what follows to stay true to what can be gleaned from the literature. I first define an interpretation as an assignment of content to the variables. I then walk through several representational principles and show how they govern this process of content assignment, resulting in an account of what makes for a permissible interpretation. From this, I then derive precisely what the content of a model is on a given interpretation. Finally, I define what it is for a model to be accurate of some situation given its content under a particular interpretation.

§2.2 Principles of Variable Selection

I begin with the variables. The values of variables represent the causal relata. It is an often-touted benefit of the causal model framework that the way that variables represent is neutral with respect to what the relata are taken to be. Variables can be taken to represent whatever is the modeler's preferred choice of causal relata – events, facts, property instantiations, etc.. As a result, the model framework bypasses the familiar debate over the nature of causal relata. But while neutral with respect to comparisons between events, facts, property instantiations, etc., the way variables represent seems to introduce new structure into our causal reasoning about actual causation, or at the very least illuminate existing structure that has otherwise been implicit. This point has not been well recognized in the literature, and merits emphasizing.

2.2.a Variables as Ranges of Alternatives

To see where the new structure comes in, consider that judgments of actual causation have the form '*c* causes *e*,' where *c* and *e* are names of token-level things, such as the occurrence of a concrete event or the instantiation of a property by some particular object. The car swerving into oncoming traffic caused the cars to crash into each other in a 7-car pile-up. Token-level things are represented in a model by values of *variables* – that is, as the kind of thing that comes alongside some specified range of alternatives. In translating natural causal talk into the framework of a causal model, we are forced to shift from naturally speaking of the causal efficacy of token-level things like property instances *considered on their own* to speaking of them *as one of a specified set of alternatives*. Hitchcock seems to recognize this point when he writes, "In using *variables* to represent causal relations, we have changed the language that we use to talk about causal relations. (Hitchcock, 2012, p. 87)"

The crucial observation is that things like the car swerving into oncoming traffic can be considered as a member of several different ranges of alternatives. One natural alternative to the car swerving into oncoming traffic is perhaps the car continuing to drive as normal. Another is that of the car swerving onto the shoulder. And taking these two together would create yet a third distinct set of alternatives for the original cause. In order to translate a claim of actual causation into a claim in terms of a model, then, a choice must be made as to how to fill in the other values of the representing variable.[1]

Obviously, not just any way of specifying alternatives to a token-level cause will produce a collection of things suitable for being represented by a variable. The process of selecting alternatives, commonly referred to as *variable selection*, is constrained by several representational principles, commonly referred to as the *principles of variable selection*. While it is widely acknowledged that there are such principles, what they are exactly has yet to be made satisfactorily explicit. I address this here.

I take there to be at least three widely presupposed such principles, albeit rarely articulated. These are what I call *exclusivity*, *exhaustivity*, and *distinctness*. I will also discuss an additional principle that may be required – what I call *intrinsicality*.

2.2.b Exclusivity

Exclusivity and exhaustivity are straightforward. Together, they ensure that variables function as partitions. Take exclusivity first. There is a non-controversial formal constraint on models that requires that a variable not take more than one value at a time. *Exclusivity* is

[1] Because of this, we might see the success, or even promise, of causal model accounts of causation as vindicating the idea of contrastive causation (Hitchcock, 1996b, 1996a; Maslen, 2004; Northcott, 2008; Schaffer, 2005, 2012). One might think that quantifying over apt models in order to get causation simpliciter would do away with this contrastivity. However, while it does away with explicit reference to contrastivity, aptness principles will still stipulate which contrast sets are appropriate for a given inquiry. Thus, causal model accounts are arguably inherently committed to a version of contrastivity. Contrastivism is discussed further below, in §4.6.c.

a representational principle that ensures this. It holds that the values of a single variable should represent mutually exclusive things, so that the manifestation of any particular thing represented by a value of this variable excludes the manifestation of any of the others.[2]

Exclusivity seems widely endorsed. For example, Hitchcock writes,

> [I]n constructing a model, it is important to choose the variables so that different values of the same variable correspond to events (or versions of events) that are incompatible on broadly logical or conceptual grounds; typically, they will represent incompatible states of a system at the same time…. (2007a, p. 502)

And Woodward explains,

> When considering the values of a single variable, we want those values to be logically exclusive, in the sense that variable *X*'s taking value *v* excludes *X*'s also taking value *v'*, where $v \neq v'$. (2016, p. 1064)

2.2.c Exhaustivity

While exclusivity ensures that a variable takes *at most* one of its values, exhaustivity ensures that a variable takes *at least* one of its values. *Exhaustivity* is the representational principle

[2] For references to exclusivity, see (Blanchard & Schaffer, 2017, p. 182; Briggs, 2012, p. 142; Hitchcock, 2004, p. 145, 2007b, p. 76, 2007a, p. 502; Pearl, 2000, p. 3; Woodward, 2003, p. 98)

that requires that a variable's values capture the entire range of alternative possibilities for whatever type of thing the variable represents.[3] This is easily and naturally achieved for variables representing events by including a value for the occurrence of the event and one for its non-occurrence. For property-instantiations of some underlying object, it requires partitioning the property space and assigning a distinct value to every portion of the partition.[4]

Exhaustivity is also widely endorsed. In his seminal text, Judea Pearl also assumes both exclusivity and exhaustivity when defining a variable. He writes, "B_i, i = 1, 2, …, n, is a set of exhaustive and mutually exclusive propositions (called a *partition* or a *variable*)…" (2000, p. 3) Several pages later, he continues,

> Likewise, each variable X can be viewed as a partition of the states of the world, since the statement X = x defines a set of *exhaustive* and *mutually exclusive* sets of states, one for each value of x. (2000, p. 9, emphasis my own)

A qualification is commonly placed on exhaustivity that restricts it to only the *serious*, *genuine*, or *relevant* possibilities for whatever type of thing the variable represents. Blanchard and Schaffer adopt such a qualification, attributing it to Hitchcock (2001, p. 287), when they require that "[t]he variables should not be allotted values that one is not willing

[3] Note that this is a different notion than that which goes by the name 'exhaustivity' in (Franklin-Hall, 2016), where a variable is "exhaustive" relative to some effect when it captures all of the possible ways that effect could have been brought about.

[4] For references to exhaustivity, see (Blanchard & Schaffer, 2017, p. 182; Briggs, 2012, p. 142; Hitchcock, 2001, p. 287; Pearl, 2000, p. 3; Woodward, 2016, p. 1064)

to take seriously (2017, p. 182).” Woodward writes, “We also want our variables to take a range of values corresponding to the full range of genuine or serious possibilities that can be exhibited by the system of interest (2016, p. 1064).”

Whether this qualified version of exhaustivity is objective or pragmatic depends on the nature of the guiding principles behind whether a possibility counts as genuine or serious. If they are even partly pragmatic, then actual causation itself will also be pragmatic – according to any SEM definition of actual causation that relies on this principle of variable selection. But it seems unlikely that a purely objective set of conditions would be forthcoming. Furthermore, some examples seem to require the qualified version of exhaustivity in order to deliver intuitively correct verdicts. This presents a problem. But it is one that my view can solve. For now, I will leave this qualification out of the principle of exhaustivity. The work that “serious possibilities” is intended to do can still be done on the view that I develop without infecting a SEM definition of actual causation with pragmatic considerations. I return to this in Chapter 4 (§4.6.d).

2.2.d Distinctness

The third principle of variable selection is *distinctness*. Distinctness holds that things which are represented by different variables should be relevantly distinct.[5] How to define precisely the relevant notion of distinctness remains an open question. Distinctness seems to be

[5] For references to distinctness, see (Blanchard & Schaffer, 2017, p. 182; Briggs, 2012, p. 142; Hitchcock, 2004, p. 146, 2007a, p. 502; Paul & Hall, 2013, p. 59)

needed here for the same reason as it is needed in a more traditional counterfactual account of causation. Lewis qualified his counterfactual account of causation as counterfactual dependence holding between *distinct* entities (Lewis, 1973c, 2000).[6] This is needed to avoid spurious causal relations popping up as the result of counterfactual dependence that holds between things that are *conceptually* related (such as an apple's being red depending on it being crimson), *mereologically* related (such as the left-hand side of the table being made of wood depending on the whole table being made of wood), or *logically* related (such as it being the case that ϕ depending on it being the case both that $\psi \rightarrow \phi$ and that ϕ). Roughly, then, the requirement is that no two values from respectively any two different variables should stand in any logical, conceptual, or mereological dependency relations with each other.

So, distinctness is needed in order to separate the wheat of causation from the chaff of mere counterfactual dependence. Its violation permits models that are misleading in that they represent dependency where intuitively there is no causation or the lack of a dependency where intuitively there is.[7]

Distinctness is related to Woodward's principle of *independent fixability* or *independent manipulability*, which is the requirement that any variable in a model should be such that it can be fixed at any of its values without forcing any other variable to take a certain value. Obviously, the kind of forcing here is non-causal. Weslake explains this requirement as its

[6] See also (Kim, 1974) for discussion.

[7] See (Woodward, 2016, p. 1060) for a useful illustration of this.

being “metaphysically possible that every proper subset of the variables in [a model] be set to every combination of their possible values by independent interventions (forthcoming, p. 15).”[8] A model whose set of variables satisfies independent fixability also satisfies distinctness.

Whatever one might call it, this requirement that I call distinctness is widely assumed. Blanchard and Schaffer require of apt models that "[t]he values of variables should not represent events that bear logical or metaphysical relations to each other (2017, p. 182).” Woodward explains,

> Other things being equal, one should exclude choices of variables which are such that certain combinations of values for those variables are assumed, as part of the set-up of the problem, to be impossible. As this example illustrates, the relevant notion of ‘impossibility’ here may include more than logical impossibility narrowly construed – it may include constraints that arise on the basis of spatio-temporal or compositional relationships. (Woodward, 2016, pp. 1063–1064)

Finally, in their discussion of interventionist counterfactuals, Ray Briggs commits themself to each of exclusivity, exhaustivity, and distinctness,

[8] See (Woodward, 2008, 2015, 2016, pp. 1063–1064). Yang (2013) also argues for the same principle. But see (Zhong 2020) for a defense of a weaker principle – one according to which two variables are sufficiently distinct just in case at least one of them can be fixed at *at least one* value such that the other variable is free to take any of its values.

> Each V∈V comes with a range R(V) of possible values – that is, answers to the question posed by the variable. I assume that the answers to any given question are *mutually exclusive* and *jointly exhaustive*, and that no answer to one question entails an answer to any other. (2012, p. 142, emphasis my own)

It might be worth flagging that distinctness and exclusivity are logically independent.[9] Exclusivity is a constraint on an individual variable and distinctness is a constraint on a set of variables. In the one direction, a model can have only exclusive variables but still fail to satisfy distinctness. Take as an example a model that has three variables, *X*, *Y*, and *Z*. Have *X* represent whether a heat source is present or not. *X* can take the value *1* if a heat source is present and *0* if one is not present. Have *Y* represent whether a match is lit or not. *Y* can take the value *1* if a match is lit and *0* if no match is lit. Finally, have *Z* represent whether a fire occurs. Each of *X*, *Y*, and *Z* are exclusive variables. For each, taking any one value precludes the taking of any other. But clearly, *X* and *Y* are non-distinct. If *Y* takes the value *1*, then *X* must take the value *1* on pain of conceptual inconsistency. But it must do so for non-causal reasons. For a match to be lit *just is* for a heat source to be present.

In the other direction, a model can satisfy distinctness yet include a variable that is non-exclusive. Take as an example a model that has only three variables, *X*, *Y*, and *Z*. Have *X* represent the material constitution of the table, *Y* represent the lighting of a match, and *Z* whether a fire breaks out. *X* can take the values *0* if the table is made of wood, *1* if the left side

[9] Thanks to Michael Strevens for drawing my attention to their independence, which is not always recognized in the literature. Hitchcock (2007a, p. 502), for example, refers to distinctness as a *corollary* of exclusivity.

of the table is made of wood, *2* if the right side of the table is made of wood, *3* if none of the table is made of wood. Variables *X*, *Y*, and *Z* are distinct. But clearly, *X* is not exclusive. If it takes the value *0* then it must also simultaneously take the values *1* and *2*.

2.2.e Intrinsicality

So far we've had constraints on *individual variables* (exclusivity and exhaustivity) and constraints on *sets of variables* (distinctness). But there also seems to be need for a constraint on *values* of variables. In particular, there seems to be a need to place an intrinsicality restriction on what values are permitted to represent. Blanchard and Schaffer write that "The values allotted should represent intrinsic characterizations (2017, p. 182)." So, if we take values to represent property-instances, then this principle would require they represent instances of only *intrinsic* properties. If we take values to represent events, then this principle would require they represent only intrinsically characterized events.

A theory of aptness that omits an intrinsicality requirement sanctions models that deliver seemingly counterintuitive verdicts. Take the following situation as an example:

> **Plato's Grief** Socrates dies and Plato grieves the death of his teacher. But Plato has no fondness for Socrates's wife, Xanthippe, and would not be dismayed to see misfortune come her way.

We could model **Plato's Grief** with two variables – one that represents Xanthippe instantiating the property of becoming a widow or not so instantiating and a second variable that represents Plato instantiating the property of grieving or not so instantiating. Given such a model and interpretation, Xanthippe becoming a widow would satisfy ***AC – relative*** as an actual cause of Plato grieving. But this isn't quite right. As stipulated, Plato doesn't grieve the fate of Xanthippe but his dear teacher Socrates. An intrinsicality requirement would rule as inapt any variable representing Xanthippe becoming a widow, and so prevent counterintuitive verdicts like these. However, I will set this principle aside along with the "serious" qualification on exhaustivity. As I will argue in Chapter 4, the view I ultimately endorse provides a satisfying error theory for cases like **Plato's Grief**.

§2.3 What a Model Says and What it Represents

2.3.a What the Variables *Say*

So, there are three principles of variable selection: exclusivity, exhaustivity, and distinctness. These principles have been defined as rules for how to construct a model given a particular situation and how to assign content to that model. Thus, the principles have been discussed as they apply to the process of model construction. If we define an interpretation as an assignment of content to the variables of a model, then an interpretation of a model will be *permissible* just in case it satisfies all the principles of variable selection. (I show why this simple definition is incomplete in Chapter 4, but will set that complication aside for now.)

However, these principles work in the other direction, as well. Taking them as given determines what is entailed by a model on an interpretation. Against a background where it's taken as given that exclusivity holds, an assignment of content implies that any two things represented by values of the same variable are mutually exclusive. Where it's taken as given that exhaustivity holds, an assignment of content implies that the range of alternatives represented by any variable is the entire range of possible alternatives for whatever that variable represents. Anything else that would be mutually exclusive with any member of that range is rendered impossible. Finally, where it's taken as given that distinctness holds, an assignment of content implies that any two things represented as values of different variables are logically, mereologically, and conceptually independent of each other.

2.3.b What the Variables *Represent*: Property Instantiations

As mentioned above, the variables of a causal model can be taken to represent whatever is the modeler's preferred choice of causal relata. This choice determines what kind of content an interpretation assigns to the variables. I will assume that causal relata are property instances or property instantiations – a particular object's instantiating some property – such as the car's being red or the desk having mass 45kg.[10] On this assumption, an assignment of content to a variable, X, will involve specifying a particular object, a, and then mapping each value of X, x_i, to a unique monadic property, F_i, being instantiated by a.[11] In general, then, an *assignment of content* comprises, for each variable in the model, a selection

[10] See (Paul, 2000) for an alternative theory of actual causation in terms of property instantiations.

[11] Note that despite the requirement that F_i be a *monadic* property, we can still use this schema to talk about n-ary relations being instantiated if we loosely construe "object" and permit a to refer to an ordered n-tuple.

of underlying object and a one-to-one mapping of a range of properties being instantiated by that object onto that variable's range of values.

In order for such an assignment, or *interpretation*, to count as *permissible*, it must satisfy exclusivity, exhaustivity, and distinctness. An interpretation of a variable, X, satisfies *exclusivity* and *exhaustivity* just in case the set of property instances, $\{F_1a, F_2a, \ldots, F_na\}$ mapped to the range of values of X, $\{x_1, x_2, \ldots, x_n\}$, is mutually exclusive and jointly exhaustive.

Exclusivity & Exhaustivity	$X = x_1 \;\rightarrow\; F_1a$ $X = x_2 \;\rightarrow\; F_2a$ $\vdots$ $X = x_n \;\rightarrow\; F_na$ … where $\{F_1a, F_2a, \ldots, F_na\}$ is mutually exclusive and jointly exhaustive.

Figure 4.

An interpretation of any two variables, $\{X, Y\}$, satisfies *distinctness* just in case for any property instantiation, F_ia, represented by a value of X, and for any property instantiation, G_ib, represented by a value of Y, F_ia and G_ib are distinct – that is, the manifestation of one is logically independent of the manifestation of the other. It may be worth flagging that this can be satisfied when $a = b$ or when $F_i = G_i$ (but not both, of course).

Distinctness	$X = x_i \;\rightarrow\; F_ia$ $Y = y_i \;\rightarrow\; G_ib$ … where $X \neq Y$ and F_ia and G_ib are distinct.

Figure 5.

2.3.c The Assignment: Setting the Initial Conditions

Given my assumption of property instances as relata, the Assignment of a model says that whatever underlying object is represented by an exogenous variable actually instantiates the property represented by its assigned value. $\mathcal{M}_1$, for example, has the Assignment *X = 1*. So, $\mathcal{M}_1$ says that whatever underlying object is represented by the variable X actually instantiates the property represented by the value 1.

2.3.d What the Equations *Represent*: Counterfactual vs. Causal Dependencies

I will now turn to the equations. While I spoke first of what the variables *said*, and then of what they *represent*, I will take the content of the equations in reverse order. As to what the equations of a model are taken to represent, there is a choice to be made. While the equations of a SEM and arrows of a DAG are universally taken to represent dependencies, the literature divides over what *kind* of dependencies. There are two principal options. The first is that an equation represents that certain complex counterfactuals are true (Blanchard & Schaffer, 2017; N. Hall, 2007; Hitchcock, 2007a, 2009; Woodward, 2003). Proponents of this view generally continue the tradition of seeking to reduce causal dependence to counterfactual dependence, although not necessarily. The second option is that an equation represents that a causal dependency holds – generally of a type-level nature (Cartwright, 2016; Gallow, forthcoming; J. Halpern & Hitchcock, 2010; Hiddleston, 2005a; Pearl, 2000). Proponents of this view take counterfactual dependence to supervene on causal dependence, which in turn may be treated as primitive or else reduced further.

While my sympathies lie with the former, I will avoid taking sides in the matter by doing the following. Since proponents of the second view take their preferred dependencies to ground the truth of the complex counterfactuals talked about by proponents of the first, they take the equations to *imply* counterfactuals even if they don't take the equations to represent them. As a result, I can streamline and simply say that equations *entail* complex counterfactuals, remaining neutral on why they do so.[12] Happily, there is fairly general consensus on what counterfactuals are *entailed* by the equations.[13] Where theory pulls apart depending on which of these two options one chooses, and so clarity demand I discuss them separately, I will refer to the *counterfactual view* and the *causal dependence view*.

2.3.e What the Equations *Say*: Entailed Counterfactuals

Entailed counterfactuals follow straightforwardly from the equations, as well as from any Boolean combination of interventions on variables, in the following way.[14] The equation $Z := max(X, Y)$, from $\mathcal{M}_1$, for example, entails the counterfactuals:

i. $(X = 1 \wedge Y = 1) \;\square\!\!\rightarrow Z = 1$

ii. $(X = 0 \wedge Y = 0) \;\square\!\!\rightarrow Z = 0$

[12] Obviously, the semantics of the entailed counterfactuals will diverge depending on which view one holds. As a further simplification, I will rely on verdicts about their *intuitive* truth or falsity. I will assume a general consensus on the intuitive truth-conditions of the counterfactuals that I discuss.

[13] This remains a somewhat open question. The original logic of causal models, found in (Galles and Pearl 1998), only covers atomic and conjunctive antecedents and consequents. This gets extended to arbitrary Boolean consequents in (Halpern 2000) and extended again to arbitrary Boolean antecedents and counterfactual consequents in (Briggs 2012). It remains to be seen whether the logic can be extended to cover counterfactuals with counterfactual antecedents.

[14] Note that a model will entail counterfactuals that have logically complex consequents as well as antecedents. But I will focus on counterfactuals with atomic consequents as a simplifying assumption.

iii. $(X = 1 \wedge Y = 0) \;\square\!\!\rightarrow Z = 1$

iv. $(X = 0 \wedge Y = 1) \;\square\!\!\rightarrow Z = 1$

Note that, syntactically speaking, counterfactuals entailed by a causal model specify that the antecedent be set by intervention. So, (i) says that if $\mathcal{M}_i$ were intervened on to set $X = 1$ and $Y = 1$, then the value of Z would be 0. This leads to interesting implications for the evaluation of these counterfactuals when the model is given an interpretation. On the interpretation, $\mathcal{I}(\mathcal{M}_1)_{EP}$, from earlier, for example, (i) – (iv) translate into the following natural language counterfactuals:

$$\mathcal{I}(\mathcal{M}_1)_{EP}: \quad X = \begin{cases} 1\; if\; Suzy\; throws\; a\; rock \\ 0\; if\; Suzy\; doesn't\; throw \end{cases}$$

$$Y = \begin{cases} 1\; if\; Billy\; throws\; a\; rock \\ 0\; if\; Billy\; doesn't\; throw \end{cases}$$

$$Z = \begin{cases} 1\; if\; the\; window\; shatters \\ 0\; if\; the\; window\; doesn't\; shatter \end{cases}$$

v. If it were the case that Suzy throws a rock and Billy throws a rock, then the window would shatter.

vi. If it were the case that Suzy doesn't throw and Billy doesn't throw, then the window would not shatter.

vii. If it were the case that Suzy throws but Billy doesn't, then the window would shatter.

viii. If it were the case that Suzy doesn't throw but Billy throws, then the window would shatter.

Carried over from the syntax is the stipulation that the antecedent be treated as being set by intervention. So, in order for (e) to be true relative to the represented situation, it is not enough that the antecedent is actually true and the consequent is actually true. It must also be the case that the consequent holds when the antecedent is set by intervention – breaking the dependence the constituents of the antecedent might have otherwise had on upstream things. This diverges from more traditional evaluations of causal counterfactuals, such as similarity semantics, on which the antecedent being true and the consequent being true in the actual situation is sufficient for the counterfactual to be true. In effect, there is a presupposition behind the use of this framework that counterfactuals entailed by a causal model are evaluated relative to a semantics that assumes at most weak centering. Menzies writes, "There is no presumption in this framework that when an antecedent is true the set of closest antecedent-worlds is restricted to the actual world (2008, p. 207)."

§2.4 Accuracy

Generally, a model on an interpretation will be accurate of some situation whenever what the model says on that interpretation is true of that situation. Now that we have a clear view of what a model says, we can make this more precise.

2.4.a Defining Accuracy

To begin, accuracy requires a permissible interpretation. I have defined a permissible interpretation as one that satisfies exclusivity, exhaustivity, and distinctness. A permissible interpretation is therefore one that says truly of two property instantiations that they are exclusive, exhaustive, or distinct just in case they really are exclusive, exhaustive, or distinct in that situation.

Accuracy is then a function of what the remaining components of the model say given a permissible interpretation. These are the Assignment and the equations. First, a model will be accurate only if the Assignment says something true. More precisely, a model, $\mathcal{M}_i$, on a given interpretation, $\mathcal{I}(\mathcal{M}_i)$, is accurate of some situation only if whatever $\boldsymbol{\mathcal{A}}_{\mathcal{M}_i}$ says is the case given $\mathcal{I}(\mathcal{M}_i)$ is indeed the case in that situation.

Finally, a model will be accurate only if the counterfactuals it entails are true. More precisely, a model, $\mathcal{M}_i$, on a given interpretation, $\mathcal{I}(\mathcal{M}_i)$, is accurate of some situation only if the counterfactuals entailed by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$ given $\mathcal{I}(\mathcal{M}_i)$ are true. Of course, whether the model is accurate *because* the counterfactuals are true – that is, whether the truth of the counterfactuals grounds the accuracy of the model – will depend on one's position in the earlier debate about what the equations represent. If one takes the counterfactual view, then a model is accurate simply in virtue of the truth of the counterfactuals. If one instead takes the causal dependence view, then a model is accurate in virtue of representing causal dependencies that actually hold in the target situation. Getting these dependencies right is then what explains why an accurate model entails true counterfactuals.

Putting this together gives the following account of what makes a model-interpretation pair *accurate* of some situation:

> **Accuracy:** A causal model, $\mathcal{M}_i$, on an interpretation $\mathcal{I}(\mathcal{M}_i)$, is accurate of a given situation, $\mathbb{S}$, just in case …
>
> i. $\mathcal{I}(\mathcal{M}_i)$ is a permissible interpretation of $\mathcal{M}_i$ for representing $\mathbb{S}$;
>
> ii. The content entailed by the assignment, $\boldsymbol{\mathcal{A}}_{\mathcal{M}_i}$, on $\mathcal{I}(\mathcal{M}_i)$ is the case in $\mathbb{S}$;
>
> iii. The counterfactuals entailed by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$ on $\mathcal{I}(\mathcal{M}_i)$ are true in $\mathbb{S}$.

2.4.b An Application

Let's see this in action with a new example. Consider the following target situation:

> **Safe Driving** Steve drives his car down the road and approaches a streetlight. The light is green, so Steve continues driving at his same speed through the light. Had the light been red, Steve would have stopped. A nearby traffic officer looks on. He would have issued a ticket had the light been red and had Steve continued through it.[15]

[15] A quick comment about how to read vignettes of this type: I take Gricean's communicative maxims to apply when confronted with vignettes or hypothetical situations – most relevantly, the maxims of Quantity and Relation (Grice, 1989). Doing so permits the assumption that all relevant and required information has been provided and nothing more. We can therefore safely fill in extraneous details in ways that are normal. So, we can assume that the traffic light is operating normally, which means that it will only shine one color at a time; that the laws of physics hold, so the car will only come to a stop if a force is acted upon it; etc.

In an effort to model **Safe Driving**, give $\mathcal{M}_1$ the following interpretation, $\mathcal{I}(\mathcal{M}_1)_{SD}$: X represents a streetlight, which takes the value 1 if the streetlight instantiates green and 0 if it instantiates red. Y represents a driver's action, which takes the value 1 if the driver instantiates stopping and 0 if he instantiates continuing at the same speed. Finally, Z represents a traffic officer's action, which takes the value 1 if he doesn't instantiate issuing a ticket and 0 if he does.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{X\}$
$\boldsymbol{V} = \{Y, Z\}$
$\boldsymbol{R} = f(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $Y := 1 - X$
(EQ3) $Z := \max(X, Y)$

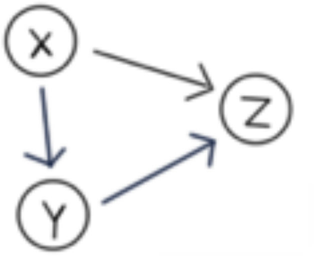


Figure 1. (again) $\boldsymbol{\mathcal{M}_1}$

$\mathcal{I}(\mathcal{M}_1)_{SD}$: *X (streetlight)* := $\begin{cases} 1\ if\ green \\ 0\ if\ red \end{cases}$

Y (driver's action) := $\begin{cases} 1\ if\ stops \\ 0\ if\ continues\ at\ same\ speed \end{cases}$

Z (traffic officer's action) := $\begin{cases} 1\ if\ doesn't\ issue\ a\ ticket \\ 0\ if\ issues\ a\ ticket \end{cases}$

We can now ask whether $\mathcal{M}_1$ on $\mathcal{I}(\mathcal{M}_1)_{SD}$ is accurate of **Safe Driving**. This amounts to asking whether (i) – (iii) are each satisfied. I will take these in turn. First, (i) is satisfied – $\mathcal{I}_{\mathcal{M}1}$is a permissible interpretation for representing **Safe Driving.** To see this, note that $\mathcal{I}(\mathcal{M}_1)_{SD}$ will be a permissible interpretation for representing **Safe Driving** if and only if whatever it says is exclusive, exhaustive, and distinct really are exclusive, exhaustive, and distinct in **Safe Driving**. Take *exclusivity* first. $\mathcal{I}(\mathcal{M}_1)_{SD}$ says that the light's being red is mutually exclusive from its being green, the driver's stopping is mutually exclusive from his continuing on at the same speed, and the traffic officer's issuing a ticket is mutually exclusive from his not doing so. These are all the case in this situation.

$\mathcal{I}(\mathcal{M}_1)_{SD}$ says, as the result of *exhaustivity*, that the light could only have been red or green, the driver could only have continued at the same speed or stopped, and the traffic officer could only have issued a ticket or not. This is also the case in this situation.

Finally, $\mathcal{I}(\mathcal{M}_1)_{SD}$ says, as the result of *distinctness*, that the light's being red or green is metaphysically distinct from the driver's continuing or stopping and from the traffic officer's issuing a ticket or not, and that the driver's continuing or stopping is metaphysically distinct from the traffic officer's issuing a ticket or not. None of these property instances entails any of the others. So far so good.

Next, (ii) is satisfied. On $\mathcal{I}(\mathcal{M}_1)_{SD}$, $\boldsymbol{\mathcal{A}}_{\mathcal{M}1}$ says that the light is green, which is true of **Safe Driving**. Finally, (iii) is satisfied. EQ2, ($Y := 1 - X$), entails the counterfactuals, (a) $X = 1 \mathrel{\square\!\!\rightarrow} Y = 0$; and (b) $X = 0 \mathrel{\square\!\!\rightarrow} Y = 1$. On $\mathcal{I}(\mathcal{M}_1)_{SD}$, these are translated to mean that (a) had the light been green, then Steve would have continued at the same speed; and (b) had the light been red, then Steve would have stopped. These will be true of the situation assuming that Steve is an alert, law-abiding driver. **Safe Driving** doesn't specify whether this assumption holds. But since the target situation is a vignette, and not a real world situation, we can safely assume they hold so long as they serve our purpose and are assumed by all interlocutors.

EQ3, $Z := \max(X, Y)$, entails the counterfactuals already indicated in §2.3.e. On $\mathcal{I}(\mathcal{M}_1)_{SD}$, these mean that (c) had the light been green and Steve stopped, then the officer would *not* have issued a ticket; (d) had the light been red and Steve continued at the same speed, then the officer *would have* issued a ticket; (e) had the light been green and Steve continued at the same speed, then the officer would *not* have issued a ticket; and (f) had the light been red and Steve had stopped, then the officer would *not* have issued a ticket. In short, the officer would issue a ticket only had the light been red and Steve continued at the same speed. This is true of the situation under the assumptions that the officer is alert, always gives tickets to anyone who runs a red light, and does not issue tickets to those who stop at green lights. Again, since **Safe Driving** is a vignette, we can safely assume these assumptions so long as all interlocutors agree. Since (i) – (iii) are each satisfied, $\mathcal{M}_1$ on $\mathcal{I}(\mathcal{M}_1)_{SD}$ is accurate of **Safe Driving**.

2.4.c Accuracy and Aptness

Is accuracy all we need for aptness? While a good start, this definition of accuracy is inadequate for aptness as it stands. There are two reasons for this. The first can be seen in the problem of structural isomorphs, a recently discovered challenge to extant SEM definitions of actual causation. I will demonstrate this problem in the next chapter, examine standing solutions to it, explain why these solutions fall short, and finally extend my own solution in the form of a new aptness condition.

The second reason for the inadequacy of this definition of accuracy is due to something that has so far gone unnoticed in the literature. In Chapter 4, I will demonstrate how a model on an interpretation under-specifies the background possibilities. Filling in these possibilities in one way results in a model on a given interpretation being accurate, while filling them in in another way results in the same model on the same interpretation being inaccurate. I argue that this calls for an enrichment of what constitutes an interpretation – that an interpretation include a specification of background possibilities. I then explore the philosophical ramifications that this enrichment has on SEM definitions of actual causation.

CHAPTER 3

Structural Isomorphs and Explicit Partial Mediation

§3.0 Abstract The recently discovered problem of structural isomorphs makes the task of articulating aptness especially salient (Blanchard & Schaffer, 2017; N. Hall, 2007; Hitchcock, 2007a; Menzies, 2017). This paper presents a new aptness requirement, *Explicit Partial Mediation*, that resolves the problem of structural isomorphs. I propose that Explicit Partial Mediation replace the aptness requirement from the literature, what I call *Essential Structure*, which enjoins us to represent enough so as to capture the target situation's essential structure.

§3.1 Introduction

The question of when a causal model is apt has become all the more pressing due to the recently discovered *problem of structural isomorphs*. [1] This problem occurs when the same model can accurately represent two different situations and yet our judgment of what causes

[1] Also called the problem of *counterfactual* isomorphs. See (Blanchard & Schaffer, 2017; N. Hall, 2007; Hiddleston, 2005b; Hitchcock, 2007a; Menzies, 2017) for additional presentations of the problem.

what differs in the two situations. Cases like this suggest that more than accuracy is needed to render a model *apt* for any given situation.

What we need, then, is some way to rule such models inapt for representing one (or both) of the situations. But the standard way to do this – by distinguishing between default and deviant states of a system – gives up on an objective account of actual causation (Gallow, forthcoming; N. Hall, 2007; J. Halpern & Hitchcock, 2010; J. Y. Halpern, 2016b; J. Y. Halpern & Hitchcock, 2015). Blanchard and Schaffer (2017) present an alternative in the way of an aptness requirement, Essential Structure, which enjoins us to include "enough events to capture the essential structure of the situation being modelled (2017, p. 183)" But this is inadequate insofar as it is left unclear what structure counts as essential and whether the underlying rationale for this requirement can be given in objective terms.

This chapter aims to rectify these inadequacies. I begin by uncovering and clarifying what underlies Essential Structure. I argue that it is the need to be sensitive to the presence of a *partially mediating structure*, which I define. Omitting such a structure prevents a model from differentiating between distinct paths between the same two variables, which undermines its ability to analyze the activity on each path separately from the other(s). By requiring that partially mediating structure be explicitly represented, the problem of structural isomorphs dissolves. I call this new aptness principle *Explicit Partial Mediation*, show how it does the work of Essential Structure, and propose that it therefore replace Essential Structure.

§3.2 The Problem of Structural Isomorphs

3.2.a Symmetric Overdetermination: a SEM in Three Parts

Consider the following situation, which is a case of symmetric overdetermination.

> **Overdetermination** Suzy and Billy each throw a brick at a window, at the same time and with the same velocity. The bricks simultaneously arrive at the window, and the window shatters.

We can represent this situation with a structural equation model (SEM). Consider a model which represents Suzy's throwing her rock or not as the two values of a binary variable. The same applies for Billy's throwing or not and with the window's being shattered or not. This results in a model with three binary variables, interpreted in the following way:

$\mathcal{I}(\mathcal{M}_2)_O$: $X\ (Suzy) = \begin{cases} 1\ if\ throws\ a\ rock \\ 0\ if\ doesn't\ throw \end{cases}$

$$Y\ (Billy) = \begin{cases} 1\ if\ throwing\ a\ rock \\ 0\ if\ not\ throwing \end{cases}$$

$$Z\ (window) = \begin{cases} 1\ if\ shattered \\ 0\ if\ not\ shattered \end{cases}$$

In this example, X and Y are the exogenous variables, Z, is the endogenous variable, and each is mapped to {*0*, *1*}. The Assignment, $\boldsymbol{\mathcal{A}}$, would be $X = 1$ and $Y = 1$, representing the fact that Suzy threw a rock and Billy threw a rock (see *Figure 6*, below).

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{X, Y\}$
$\boldsymbol{V} = \{Z\}$
$\boldsymbol{R} = f\,(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$
(EQ2) $Y = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ3) $Z := \max(X, Y)$

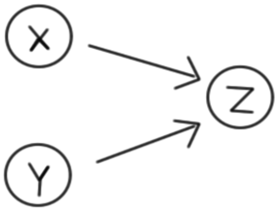


Figure 6. $\boldsymbol{\mathcal{M}_2}$

Finally, our model needs to represent the dependence of the window shattering on the children's throws with a functional equation whose form captures what actually happens in the case as well as what would have happened had the alternatives occurred instead. In this case, had either child thrown a rock the window would have shattered, and had neither child thrown it would not have shattered. This can be captured by the functional equation, *Z := max(X, Y)*.

The content of EQ3 from $\mathcal{M}_2$ can be unpacked in the following counterfactuals:

$X = 1$ (Suzy throwing) □→ $Z = 1$ (window shattered).

$Y = 1$ (Billy throwing) □→ $Z =1$ (window shattered).

$X = 0$ and $Y = 0$ (Suzy not throwing and Billy not throwing) □→ $Z = 0$ (window not shattered).

3.2.b The Problem of Structural Isomorphs

The problem of *structural isomorphs* occurs when one model can accurately represent two different situations – under two different interpretations, of course – and so delivers the same verdicts in both for what causes what. And yet our judgment of what causes what differs in the two situations. As an example, consider the following:

> **Bogus Prevention** An assassin intends to put poison in the King's coffee, but at the very last-minute changes her mind and refrains. The King's bodyguard, though, compulsively puts antidote in the King's coffee every morning, and had already done so this morning. The King survives.[2]

Compare **Bogus Prevention** with **Overdetermination**. Intuitively, the bodyguard's administration of antidote is not a cause of the King surviving. The King would have survived regardless of his bodyguard's compulsive action. And while intuitions about symmetric overdetermination cases like **Overdetermination** aren't as vivid, it is natural enough to take Suzy's throwing and Billy's throwing to be equal and independent actual causes of the window being shattered. This means that Billy's throwing is causally disanalogous to the bodyguard's administering antidote. The problem is that **Bogus Prevention** can be represented accurately using the very same model, $\mathcal{M}_2$, as was constructed to represent **Overdetermination**. We just need to interpret it differently. Use $\mathcal{I}(\mathcal{M}_2)_{BP}$ to represent **Bogus Prevention**:

[2] This example attributed to (Blanchard & Schaffer, 2017, p. 185; Hitchcock, 2007a)

$$\mathcal{I}(\mathcal{M}_2)_{BP}: \qquad X\,(assassin) := \begin{cases} 1 \; if \; not \; administering \; poison \\ 0 \; if \; administering \; poison \end{cases}$$

$$Y\,(bodyguard) := \begin{cases} 1 \; if \; administering \; antidote \\ 0 \; if \; not \; administering \; antidote \end{cases}$$

$$Z\,(King) := \begin{cases} 1 \; if \; survives \\ 0 \; if \; dies \end{cases}$$

$\mathcal{M}_2$ is accurate for both **Overdetermination** on $\mathcal{I}(\mathcal{M}_2)_O$ and **Bogus Prevention** on$\mathcal{I}(\mathcal{M}_2)_{BP}$. To see this, note that the Assignment says truly of **Overdetermination** on $\mathcal{I}(\mathcal{M}_2)_O$ that Suzy throws and Billy throws, and says truly of **Bogus Prevention** on $\mathcal{I}(\mathcal{M}_2)_{BP}$ that the assassin does not administer poison and the bodyguard does administer antidote. Second, the counterfactuals entailed by $\mathcal{M}_2$ on $\mathcal{I}(\mathcal{M}_2)_O$ are true of **Overdetermination**, and those entailed by $\mathcal{M}_2$ on $\mathcal{I}(\mathcal{M}_2)_{BP}$ are true of **Bogus Prevention**. It is true of **Overdetermination** that had Suzy thrown, then the window would have shattered; had Billy thrown, then the window would have shattered; and had neither Suzy nor Billy thrown, then the window would not have shattered. Similarly, it is true of **Bogus Prevention** that had the assassin not administered poison, the King would have survived; had the bodyguard administered antidote, then the King would have survived; and had the assassin administered poison and the bodyguard not administered antidote, then the King would have died.

The crucial thing to note is that on the respective interpretations in question, Billy's throw is structurally analogous to the bodyguard's administration of antidote; *Y = 1* in each case. As a result, if our recipe for actual causation as applied to $\mathcal{M}_2$ delivers the right results in one case then it ipso facto delivers the wrong results in the other. The dilemma is that either our intuition is mistaken that these cases have a different causal structure, or else $\mathcal{M}_2$ is not suitable for representing one or both cases on the respective interpretation. At least one model-interpretation must be inapt, despite its accuracy.

3.2.c Defaults and Deviants vs. Essential Structure

In response to this dilemma, the literature has uniformly defended the second horn. The divisive question is then: what else beyond accuracy does aptness require? The leading response introduces a normative parameter into the causal model framework, most commonly precisified as a distinction between default and deviant states of a system (Gallow, forthcoming; N. Hall, 2007; J. Halpern & Hitchcock, 2010; J. Y. Halpern, 2016b; J. Y. Halpern & Hitchcock, 2015; Menzies, 2017). There are a couple of reasons to resist this move, however. I won't go into detail, since they are thoroughly discussed in (Blanchard & Schaffer, 2017). But suffice it here to say that this move threatens to undermine the objectivity of actual causation yet arguably fails to produce even a psychologically plausible account of actual causation.

The only alternative response on offer is presented by Blanchard and Schaffer (2017). They argue that the model can be ruled inapt for representing **Bogus Prevention** by endorsing a requirement on aptness that I will call Essential Structure. This requires of an apt model that it represent enough of a situation so as to capture its essential structure.[3] Invoking this

[3] Essential structure is originally introduced in (Hitchcock, 2007a).

principle, Blanchard and Schaffer argue that $\mathcal{M}_2$ on $\mathcal{I}(\mathcal{M}_2)_{BP}$ is inapt for representing **Bogus Prevention** because it leaves out essential structure. In particular, it leaves out the bit where the antidote neutralizes (or doesn't neutralize) the poison. Including a variable to represent this produces an apt model, $\mathcal{M}_{2+}$, that delivers the intuitively correct causal verdict.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{X, Y\}$
$\boldsymbol{V} = \{N, Z\}$
$\boldsymbol{R} = f(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$
(EQ2) $Y = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ3) $N := \min(1 - X, Y)$
(EQ4) $Z := \max(X, N)$

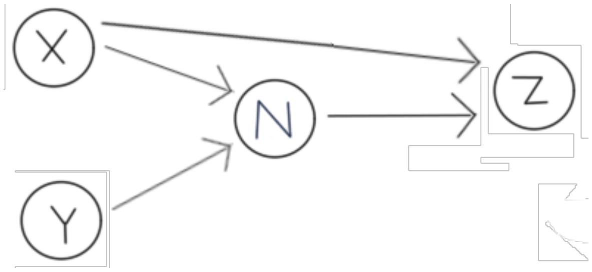


Figure 7. $\boldsymbol{\mathcal{M}_{2+}}$

$$\mathcal{I}(\mathcal{M}_{2+})_{BP}: \quad X\ (assassin) := \begin{cases} 1\ if\ not\ administering\ poison \\ 0\ if\ administering\ poison \end{cases}$$

$$Y\ (bodyguard) := \begin{cases} 1\ if\ administering\ antidote \\ 0\ if\ not\ administering\ antidote \end{cases}$$

$$N\ (neutralization\ process) := \begin{cases} 1\ if\ occuring \\ 0\ if\ not\ occuring \end{cases}$$

$$Z\ (King) := \begin{cases} 1\ if\ survives \\ 0\ if\ dies \end{cases}$$

According to $\boldsymbol{AC - relative}$**,** the new model, $\mathcal{M}_{2+}$, is not a witness to the bodyguard's administration of antidote being an actual cause of the King surviving. This is because the introduction of N renders $Y = 1$ a non-cause of $Z = 1$ in $\mathcal{M}_{2+}$. The only possible path between Y and Z is $\{Y, N, Z\}$, and there is no setting of values of off-path variables, $\{X\}$, that satisfies $\boldsymbol{AC}$2a – c. When $X = 1$, $\boldsymbol{AC}$2c isn't satisfied, and when $X = 0$, $\boldsymbol{AC}$2a isn't satisfied. Thus, $Y = 1$ is not an actual cause of $Z = 1$ relative to $\mathcal{M}_{2+}$.

§3.3 Explicit Partial Mediation

I propose an improvement on this response to the problem of structural isomorphs, which replaces Essential Structure with a new aptness requirement that I call *Explicit Partial Mediation*. Essential Structure as it stands is inadequate as an objective condition on apt causal models. Without further specification, Essential Structure remains opaque, reliant on our pre-theoretic causal intuition, and unilluminating of the nature of causation. We need an independent story about what kind of structure is essential and why. In fairness, Blanchard and Schaffer themselves concede that this doesn't yet get to the bottom of things. They write,

> [W]e think that there is a core phenomenon of *an impoverished model that omits crucial information*. We think that there needs to be some constraint corresponding to the vague idea of 'don't use impoverished models'. Any such constraint should equally be able to do the work we put [Essential Structure] … towards. (2017, n. 13)

In this section, I argue that the needed constraint is my requirement of *Explicit Partial Mediation* – that all partially mediating variables be explicitly represented – which I define shortly. I therefore propose that Explicit Partial Mediation replace Essential Structure. On the new view, a model is impoverished, in Blanchard and Schaffer's sense, insofar as it omits partially mediating variables, and enriching the model in the relevant sense is just to include them. I defend this proposal by demonstrating how this requirement does the work of Essential Structure, on the one hand, and showing how the need for the required structure is independently motivated, on the other. The account of aptness that results from replacing Essential Structure with Explicit Partial Mediation does not invoke causal notions, and so protects reductive aspirations. But it should be illuminating for non-reductionists, as well.

3.3.a Fully Mediating and Partially Mediating Variables

Interpolating phenomena are things which occur as an intermediate step in the chain of dependence between a cause and an effect. An interpolating *variable* represents some such interpolating phenomenon intermediate between an explicitly represented cause and effect in a model. A variable, Y, in a given model, $\mathcal{M}_i$, *interpolates* between two other variables (or sets of variables) in $\mathcal{M}_i$, X and Z, just in case Y figures in the X-equation and Z figures in the Y-equation. In words, Y interpolates between X and Z whenever Z depends on Y, which in turn depends on X. But there are two different kinds of interpolating variables: those that *fully mediate* between their flanking variables and those that only *partially mediate* between their flanking variables.

A fully mediating variable can be helpfully illustrated using **Overdetermination**. There are countless phenomena intermediate between Suzy throwing her brick and the window shattering. Take the phenomenon of the brick being at the midway point between her hand

and the window. Although we don't explicitly represent this, we assume that whenever the model sets X = 1, it implicitly represents the brick passing through the midway point. In addition, we assume that whenever the model sets X = 0, it implicitly represents the brick *not* passing through the midway point. A variable can be introduced into the model to explicitly capture this phenomenon. Say we add W: {1, 0} to $\mathcal{M}_1$, and add to $\mathcal{I}_{\mathcal{M}2}$ the interpretive assignment of W = 1 as the brick's being at the midway point and W = 0 as the brick's not being at the midway point. This produces the following amended model, $\mathcal{M}_{2INT}$:

$\boldsymbol{\mathcal{S}}$ = $\mathbf{U} = \{X, Y\}$
$\mathbf{V} = \{W, Z\}$
$\mathbf{R} = f(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$
(EQ2) $Y = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ3) $N := \min(1 - X, Y)$
(EQ4) $Z := \max(X, N)$

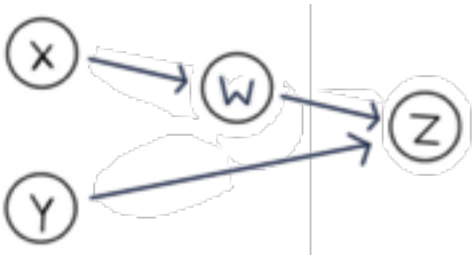


Figure 8. $\boldsymbol{\mathcal{M}_{2INT}}$

Notice in the corresponding DAG that W cuts in fully between X and Z. Call a variable like W a *fully mediating* variable. To use the terminology of screening off, a fully mediating variable screens its child variables off from its parent variables. This means that once W is held fixed at a value, no variation on X will result in any variation in Z. A variable, W, fully mediates between two variables (or sets of variables), X and Z, just in case X figures in the W-equation and W replaces X in the Z-equation. That W fully mediates between X and Z in $\mathcal{I}_{\mathcal{M}1INT}$ means that any dependence that the window shattering (or not) has on Suzy's throwing (or not) it has *fully in virtue of* the phenomenon of the brick passing (or not) through the midway point.

A helpful illustration of a partially mediating variable comes from **Bogus Prevention**. There are countless phenomena between the assassin not administering the poison and the King surviving. Take the phenomenon of there being no neutralization of poison by the antidote. As before, a variable can be introduced into the model to explicitly capture this phenomenon. Adding the variable, N: {1, 0}, to $\mathcal{M}_2$ to produce $\mathcal{M}_{2+}$, and interpret N = 1 as there being a neutralization of poison, and N = 0 as there not being a neutralization of poison.

What's crucial to recognize in the DAG corresponding to $\mathcal{M}_{2+}$ is that N does not cut in fully between X and Z. Call a variable like N a *partially mediating* variable. To use the terminology of screening off, a partially mediating variable does not screen its child variables off from its parent variables. There exists a value of N (namely, *N =0)* such that, were N held fixed at that value, variation in X may result in some variation in Z. A variable, N, partially mediates between two variables (or sets of variables), X and Z, just in case X figures in the N-equation and both X and N figure in the Z-equation. Although N mediates between its flanking variables, X and Z, there is also an independent path from X to Z, not mediated by N.

3.3.b A New Aptness Principle: Explicit Partial Mediation

I claim that while a model may benignly omit a fully mediating variable and still aptly capture the causal structure of a situation, omission of a partially mediating variable will produce an inapt model. Indeed, the omission of uncountably many fully mediating variables is formally necessitated by the discrete nature of a finite SEM coupled with the presumably dense nature of reality. But to omit a partially mediating variable is to collapse two distinct paths of dependence into one. As discussed in Chapter 1, SEM definitions of actual causation rely on the distinction between variables that are on-path and those that are off-path. Indeed, this distinction is what allows for their trademark solutions to redundant causation, as I've illustrated. It is unsurprising that the success of such a definition will be compromised when our representation conflates distinct paths. $\mathcal{M}_2$ is inapt for representing **Bogus Prevention** due to its omission of the partially mediating structure represented by N. Through this omission, the model conflates distinct paths of influence, and thereby fails to adequately capture the causal structure of its target situation. The explicit inclusion of this structure is what makes $\mathcal{M}_{2+}$ apt.

In line with this observation, I propose the following aptness requirement.

> **Explicit Partial Mediation** Include all variables that partially mediate between any two variables in the model.[4]

Explicit Partial Mediation improves on Essential Structure in specifying the kind of structure a model must include – partially mediating structure. It is a further improvement in that the need to include partially mediating structure is independently motivated. It derives from the previously discussed reason for the success of causal models: their ability to distinguish between two separate causal paths between the same two things.

[4] Hiddleston (2005a, p. 649) seems to suggest something along the lines of this proposal, although his suggestion is too spare to know for sure. This proposal can also be seen as one way of precisifying an idea of Halpern and Hitchcock, who propose that only when the addition of variables to a model changes its "topology" will those additions affect the relations of actual causation (J. Halpern & Hitchcock, 2010, p. 395).

Finally, it is worth noting that Explicit Partial Mediation will resolve Hall's (2007) original concern which launched the discussion of structural isomorphs. Hall argues that a SEM definition of actual causation will always mistakenly ascribe causation to any preventative measure regardless of whether it actively protects against a live threat or merely safeguards against possible but non-actual threats. This is the difference between the neighborhood patrol stopping a burglary from taking place, and so causing the family to sleep peacefully through the night, and the neighborhood patrol merely safeguarding the family from any possible burglaries, although none are actually attempted. In the first instance the neighborhood patrol causes the family to sleep peacefully through the night by actively preventing the burglary. In the second instance, the neighborhood patrol does not cause the family to sleep peacefully, although it would have had a burglary been attempted. Note that the bodyguard's administration of antidote is just such a safeguard, too, in a situation without poison such as **Bogus Prevention**.

So, there is a real causal difference between active protectors and mere safeguards. Hall insists that we need a default/deviant distinction in order to distinguish them. Proponents of Essential Structure can instead say that some essential structure must be missing – although without identifying what it is. But Explicit Partial Mediation can fill in the details in purely structural terms. For any model that accurately represents a situation where the preventative measure is merely a safeguard, and which explicitly represents the threat against which the safeguard protects, there may always be introduced a variable to represent whether the prevention – whatever it is – occurs or not. The introduced variable will partially mediate between the (non-existent) threat and the safeguard, on the one side, and the effect in question, on the other. Once it is explicitly included, the safeguard no longer satisfies $\boldsymbol{AC - relative}$ as a cause of the effect in question. Explicit Partial Mediation gives us a way to distinguish between active protectors and mere safeguards without relying on a normality parameter.

3.3.c Bogus Antidote

Explicit Partial Mediation does the work of Essential Structure in another case. This is the structural isomorphism that can be made to hold between **Bogus Antidote** and **Early Preemption**, which I will explain shortly. I should foreshadow, though, that while Explicit Partial Mediation dissolves the isomorphism, it does not solve everything that might seem wrong in this case. Indeed, neither does Essential Structure. I mention it here just to show that the essential structure that gets introduced is partially mediating structure, and so Explicit Partial Mediation can replace Essential Structure.

Consider first the following situation:

> **Bogus Antidote** The King's bodyguard accidentally spills some antidote into the King's coffee. The assassin sees this. She has an obligation to poison the coffee, but doesn't want to actually kill the King. Now, she can poison the coffee without risking killing the King. She does so. The King drinks the coffee and survives. (Blanchard & Schaffer, 2017, p. 202)

This is an instance of what Hall calls a “short circuit” (2007, p. 120). Intuitively, the bodyguard’s putting the antidote in the coffee is not a cause of the King surviving. It is true that the antidote prevents the poison from killing the King, but the only reason for there being poison in the first place is the presence of the antidote. The only threat subdued by the antidote is one that it produces.

The problem is that **Bogus Antidote** is structurally isomorphic to a case of early preemption. Consider:

> **Early Preemption** Suzy throws a rock at a window, the rock hits the window, and the window shatters. Her friend Billy stands by. Had Suzy not thrown, then Billy would have. And had Billy thrown, the window would still have shattered.

Here, Suzy’s throw causes the window to shatter, despite the fact that had Suzy not thrown, then Billy would have, and the window still would have shattered. So, Suzy’s throw in **Early Preemption** is intuitively disanalogous to the bodyguard’s administration of antidote in **Bogus Antidote**. Yet the same model, $\mathcal{M}_1$, can be interpreted so as to accurately represent both situations.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{X\}$
$\boldsymbol{V} = \{Y, Z\}$
$\boldsymbol{R} = f\,(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $Y := 1 - X$
(EQ3) $Z := \max\,(X, Y)$

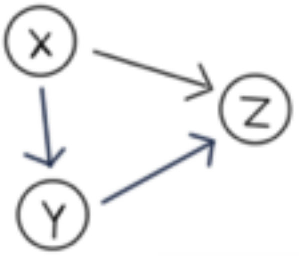


Figure 1. (again) $\boldsymbol{\mathcal{M}_1}$

Use $\mathcal{I}(\mathcal{M}_1)_{EP}$ to represent **Early Preemption**:

$\mathcal{I}(\mathcal{M}_1)_{EP}$: $X\ (Suzy) = \begin{cases} 1\ if\ throwing\ a\ rock \\ 0\ if\ not\ throwing \end{cases}$

$Y\ (Billy)\ = \begin{cases} 1\ if\ throwing\ a\ rock \\ 0\ if\ not\ throwing \end{cases}$

$Z\ (window)\ = \begin{cases} 1\ if\ shattered \\ 0\ if\ not\ shattered \end{cases}$

And use $\mathcal{I}(\mathcal{M}_1)_{BA}$ to represent **Bogus Antidote**:

$\mathcal{I}(\mathcal{M}_1)_{BA}$: $X\ (bodyguard)\ := \begin{cases} 1\ if\ administering\ antidote \\ 0\ if\ not\ administering\ antidote \end{cases}$

$Y\ (assassin)\ := \begin{cases} 1\ if\ not\ administering\ poison \\ 0\ if\ administering\ poison \end{cases}$

$Z\ (King)\ := \begin{cases} 1\ if\ survives \\ 0\ if\ dies \end{cases}$

When $\boldsymbol{AC - relative}$ is applied to $\mathcal{M}_1$, it says that *X = 1* is an actual cause of *Z = 1*. The path, {*X*, *Z*} is such that when off-path variables (*Y*) are held fixed at their actual values (*Y = 0*), then if it were the case that *X = 1* then *Z = 1*, and if it were the case that *X = 0* then *Z = 0*. This is the result we want for **Early Preemption**. *X = 1* represents Suzy throwing, which is indeed the intuitive cause of the window being shattered. However, this is not the result we want for **Bogus Antidote**. *X = 1* represents the bodyguard administering antidote, which is intuitively *not* an actual cause of the King surviving.

Blanchard and Schaffer (2017) use Essential Structure to respond to this problem, enriching the model so as to adequately capture **Bogus Antidote**. $\mathcal{M}_2$ is impoverished in leaving out essential structure, namely whether there is neutralization of poison or not. Representing this feature of the situation, using variable *N*, produces $\mathcal{M}_{2+}$.

$\boldsymbol{\mathcal{S}}$ = **U** = {*X*}
V = {*Y*, *N*, *Z*}
R = $f(X_i)$ = {1, 0}

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $Y := (1 - X)$
(EQ3) $N := \min(X, (1 - Y))$
(EQ4) $Z := \max(N, Y)$

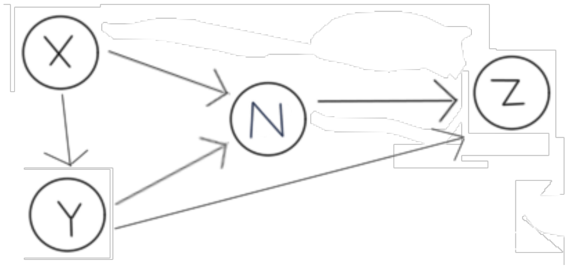


Figure 9. $\boldsymbol{\mathcal{M}_{1+}}$

By requiring the inclusion of *N*, the isomorphism between **Bogus Antidote** and **Early Preemption** is broken. But *N* is a partially mediating variable. So, Explicit Partial Mediation motivates the same response. $\mathcal{M}_1$ is ruled inapt for representing **Bogus Antidote** due to its omission of this partially mediating structure.

Of course, as mentioned above, $\boldsymbol{AC - relative}$ still gets the wrong result in the enriched model. Even with *N* included, the bodyguard's administration of antidote remains a cause of the King surviving – that is, *X = 1 is* a cause of *Z = 1*. The relevant path between *X* and *Z* is {*X, N, Z*}, and the relevant setting of values of off-path variables, {*Y*}, is the actual value, *Y = 0*.

I will say first that I am not committed to this being a problem. I am open to accepting the seemingly counterintuitive result that the antidote *does* cause the King's survival. In these kinds of short circuit cases, I don't have strong intuitions to begin with. Perhaps our resistance to the causal claim that the administration of antidote causes the King to survive stems from the role causal judgments play in responsibility ascriptions. This claim on its own would normally result in approval for the bodyguard. However, the full story negates the bodyguard's good standing by showing that the only threat to the King's life that the bodyguard's action protected against it also caused.

But Blanchard and Schaffer take this to indicate that $\boldsymbol{AC - relative}$ is inadequate and needs to be adjusted. In other words, this is a problem for the recipe of actual causation to solve. Having demonstrated that Explicit Partial Mediation does the work of Essential Structure, I could adopt this line, as well.

§3.4 Explicit Partial Mediation and Stability

I have so far argued that sensitivity to the presence (or absence) of a partially mediating structure distinguishes between the putatively structurally isomorphic situations in problem cases. Explicit Partial Mediation therefore does the work of Essential Structure and can effectively replace it as an aptness condition. Is Explicit Partial Mediation all we need, then?[5] Another oft-mentioned requirement, indeed one which Blanchard and Schaffer take to be a natural partner of Essential Structure, is Stability. Stability places a condition on an apt model that merely "[a]dding additional variables should not overturn the causal verdicts (Blanchard & Schaffer, 2017, p. 183)."[6] Does Explicit Partial Mediation obviate the need for Stability, or is Stability still needed as an additional requirement? I first demonstrate the ways in which Explicit Partial Mediation does some of the work done by Stability. However, there seem to be unstable models that are not due to the omission of partially mediating structure, and which therefore cannot be explained by Explicit Partial Mediation. After illustrating this, I argue that such cases in fact fail to show that Explicit Partial Mediation is inadequate. I conclude that Stability is not, in the end, needed.[7]

3.4.a Explained Instability

We have already seen how Explicit Partial Mediation can explain what's wrong with an unstable model. For example, $\mathcal{M}_1$ on $\mathcal{I}(\mathcal{M}_1)_{BP}$ as a representation of **Bogus Prevention** says that the bodyguard's administration of antidote is an actual cause of the King surviving. But enrich the model to include N and the new model-interpretation pair, $< \mathcal{M}_{1+},\ \mathcal{I}(\mathcal{M}_1)_{BP+} >$, says that the bodyguard's administration of antidote is *not* an actual cause of the King

[5] Once again, bracketing issues to do with what variable selection. See fn. 4, above.

[6] See also (J. Halpern & Hitchcock, 2010; J. Y. Halpern, 2016b). Note that this is a distinct notion from that termed 'Stability' and discussed by Woodward (Woodward, 2006, 2010, 2016, 2018) in the context of causal explanation.

[7] This of course leaves open the question of whether Stability is a good heuristic to apply when engaged in *causal discovery*, which is a notably different application of the SEM framework.

surviving. Explicit Partial Mediation explains that this instability as due to $< \mathcal{M}_1,\ \mathcal{I}(\mathcal{M}_1)_{BP} >$ omitting partially mediating structure, when thus rules it inapt.

Explicit Partial Mediation also explains what's wrong with another well-known case of instability, Halpern's infinite series of alternating models (J. Y. Halpern, 2016b, sec. 6). In his Theorem 6.1, Halpern proves the existence of an ordered series of models, the even members of which say that *A = 1* is an actual cause of *B = 1*, and the odd members of which say that *A = 1* is *not* an actual cause of *B = 1.* Halpern stops this infinite alternation with a normality parameter. However, Explicit Partial Mediation does the job on its own. It can be observed that every odd-membered model in this series omits a partially mediating variable, thus violating Explicit Partial Mediation. To demonstrate, I will focus on the first iteration of the series. This will be sufficient for my purposes since the series is defined inductively. I also relegate the details of the functional equations to footnotes, since they aren't needed for the argument. We begin with $\mathcal{M}_{H-0}$:

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U}$ = {U}
$\boldsymbol{V}$ = {A, B}
$\boldsymbol{R}$ = $f\,(X_i)$ = {1, 0}

$\boldsymbol{\mathcal{A}}$ = (EQ1) U = 1

$\boldsymbol{\mathcal{L}}$ = (EQ2) $A := U$
(EQ3) $B := U$

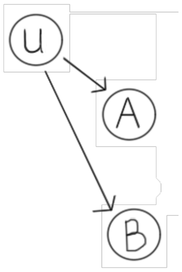


Figure 10. $\boldsymbol{\mathcal{M}_{H-0}}$

We can see that in $\mathcal{M}_{H-0}$, $A = 1$ is not an actual cause of $B = 1$. We then add variable, X_1, to create $\mathcal{M}_{H-1}$:

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U}$ = {U}
$\boldsymbol{V}$ = {A, B, X_1}
$\boldsymbol{R}$ = f (X_i) = {1, 0}

$\boldsymbol{\mathcal{A}}$ = (EQ1) $U = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $A := U$
(EQ3) $X_1 := U$
(EQ4) $B := f_B\,(U, A, X_1)$

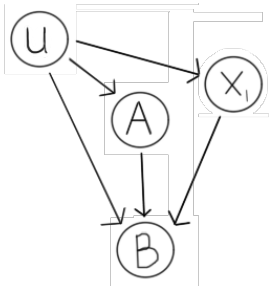


Figure 11. $\boldsymbol{\mathcal{M}_{H-1}}$

The causal verdict delivered by $\mathcal{M}_{H-1}$ is that $A = 1$ *is* an actual cause of $B = 1$.[8] We then add variable, Y_1, to create $\mathcal{M}_{H-2}$:

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U}$ = {U}
$\boldsymbol{V}$ = {A, B, X_1, Y_1}
$\boldsymbol{R}$ = f (X_i) = {1, 0}

$\boldsymbol{\mathcal{A}}$ = (EQ1) $U = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $A := U$
(EQ3) $X_1 := U$
(EQ4) $Y_1 := X_1$
(EQ5) $B := f_B\,(U, A, X_1, Y_1)$

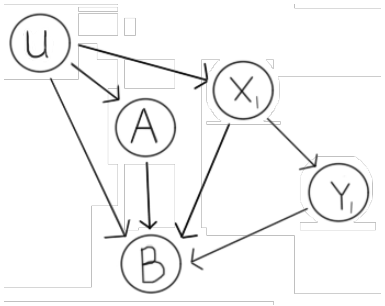


Figure 12. $\boldsymbol{\mathcal{M}_{H-2}}$

[8] To see why $A = 1$ is a cause of $B = 1$ in $\mathcal{M}_{H-1}$, the details of (EQ4) need to be given. (EQ4) takes the following form in $\mathcal{M}_{H-(2n-1)}$: $B := U$ unless $U = 1$ and either (i) $A = 0$ and either $X_n = 0$ or $X_j = Y_j = 0$ for some $j<n$; or (ii) $A = 1$ and $X_j \neq Y_j$ for some $j<n$. In $\mathcal{M}_{H-1}$, $A = 1$ satisfies $\boldsymbol{AC - relative}$ when we set $X_1 = 0$. See (J. Y. Halpern, 2016b) for further details.

The causal verdict delivered by $\mathcal{M}_{H-2}$ is that *A = 1* is *not* an actual cause of *B = 1*.[9] This alternation continues with each iterative addition of X_n and Y_n. However, notice that $\mathcal{M}_{H-1}$ omits a partially mediating variable – namely, Y_1, that mediates between X_1 and *B*. $\mathcal{M}_{H-1}$ is therefore inapt due to violating Explicit Partial Mediation. Indeed, any odd-numbered model in this series, $\mathcal{M}_{H-(2n+1)}$, will omit a partially mediating variable Y_{n+1}, and thus be ruled inapt.

3.4.b Unexplained Instability

There exist other unstable models, however, whose instability is not explained by Explicit Partial Mediation. I can demonstrate this with the earlier situation of **Overdetermination**, where Suzy's and Billy's throws simultaneously cause the window to shatter. Suppose we model **Overdetermination** with the following model and interpretation.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{X, Z\}$
$\boldsymbol{V} = \{\emptyset\}$
$\boldsymbol{R} = f(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$
(EQ2) $Z = 1$

Figure 13. $\boldsymbol{\mathcal{M}_3}$

$$\mathcal{I}(\mathcal{M}_3)_O: \qquad X\,(Suzy) = \begin{cases} 1 \; if \; throwing \; a \; rock \\ 0 \; if \; not \; throwing \end{cases}$$

[9] To see why *A = 1* is not a cause of *B = 1* in $\mathcal{M}_{H-2}$, the details of (EQ5) need to be given. (EQ5) takes the following form in $\mathcal{M}_{H-(2n)}$: *B := U* unless either (i) *A = 0* and $X_j = Y_j = 0$ for some *j<n*; or (ii) *A = 1* and $X_j \neq Y_j$ for some *j<n*. In $\mathcal{M}_{H-2}$, there is no path relative to which *A = 1* satisfies $\boldsymbol{AC-relative}$. See (J. Y. Halpern, 2016b) for further details.

$$Z\,(window) \;= \begin{cases} 1 \; if \; shattered \\ 0 \; if \; not \; shattered \end{cases}$$

Given $\mathcal{I}(\mathcal{M}_3)_O$ and with **Overdetermination** as a target, $\mathcal{M}_3$ is unstable. $< \mathcal{M}_3,\ \mathcal{I}(\mathcal{M}_3)_O >$ says that Suzy throwing is *not* a cause of the window being shattered. But enrich the model to include Billy throwing or not and Suzy throwing becomes a cause. Thus, the original causal verdict gets overturned. Yet there is no missing partially mediating structure here. The inclusion of Billy throwing is not motivated by nor does it reveal partially mediating structure. So, we have a case of instability that is not due to the omission of partially mediating structure, and thus not due to the violation of Explicit Partial Mediation.

It is clear enough that this kind of instability can be easily generated. It is not surprising that a model can fail to deliver a verdict of causation as a result of leaving too much out, yet an enriched model deliver the verdict that actual causation holds.

But note that this kind of case also does not present a challenge to any positive verdicts delivered by $\boldsymbol{AC - simpliciter}$. $\boldsymbol{AC - simpliciter}$ existentially quantifies over all apt model-interpretation pairs. So the existence of an apt model-interpretation pair that says there is *not* causation between two things will not overturn any positive verdicts by $\boldsymbol{AC - simpliciter}$ so long as there is another apt model-interpretation pair which says there *is* causation there. It is not a problem that $< \mathcal{M}_3,\ \mathcal{I}(\mathcal{M}_3)_O >$ says that Suzy's throw is *not* an actual cause of the window shattering, so long as there is another apt model-interpretation pair that says it is – and in fact, there is. Our model from earlier, $\mathcal{M}_2$, on $\mathcal{I}(\mathcal{M}_2)_O$ witnesses

Suzy's throw as a cause of the window shattering in **Overdetermination**. So the overall verdict delivered by $\boldsymbol{AC - simpliciter}$ is itself perfectly stable.

The crucial point here is that, for any definition of actual causation that existentially quantifies over a set of models, there will be a distinction between *model-relative* causal verdicts and what might be called *global* causal verdicts. Model-relative verdicts are those that result from applying the recipe of actual causation, $\boldsymbol{AC - relative}$, to an apt model-interpretation pair. Global causal verdicts are the verdicts simpliciter that result from existentially quantifying over all such applications.[10] So far we have been taking *stability* and *instability* to be local properties of particular model-interpretation pairs. A model, $\mathcal{M}_i$, is *unstable* insofar as there is at least one causal claim that it entails, of the form $X = 1$ is (is not) a cause of $Z = 1$, such that the mere inclusion of an additional variable produces a richer model, $\mathcal{M}_{i+}$, which entails the opposite causal claim, of the form $X = 1$ is not (is) a cause of $Z = 1$.[11] A model is *stable* insofar as there is no causal claim that it entails which is overturned in this way by a richer model. However, because of the existential quantification in $\boldsymbol{AC - simpliciter}$, the instability of model-relative verdicts does not necessarily translate into any instability of global verdicts.

[10] Some, such as Hall (2007; 2006), assume a definition that universally quantifies. But universal quantification – indeed any quantification – would still call for a disambiguation between model-relative verdicts and global verdicts, although the relationship between them would be of a different form.

[11] Note that I use 'mere inclusion' and 'strictly richer' as technical terms, related to Halpern's notion of *conservative extension* (J. Y. Halpern, 2016b, p. 88). Intuitively, a causal model, $\mathcal{M}_{i+}$, is a *conservative extension* of $\mathcal{M}_i$when $\mathcal{M}_{i+}$ includes an additional variable (or set of variables) but does not violate any of the dependency relations already represented in $\mathcal{M}_i$. Formally, a causal model, $\mathcal{M}_{i+}$ = ($\boldsymbol{U'}$, $\boldsymbol{V'}$, $\boldsymbol{R'}$, $\boldsymbol{A'}$, $\boldsymbol{L'}$) is a *conservative extension* of $\mathcal{M}_i$ = ($\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{R}$, $\boldsymbol{A}$, $\boldsymbol{L}$) just in case (i) $\boldsymbol{U} = \boldsymbol{U'}$; (ii) $\boldsymbol{V} \supseteq \boldsymbol{V'}$; (iii) $\boldsymbol{A} = \boldsymbol{A'}$, and (iv) for any variable, $V \in (\boldsymbol{U} \cup \boldsymbol{V})$ on any assignment of values to $X \in [(\boldsymbol{U}_{\mathcal{M}_i} \cup \boldsymbol{V}_{\mathcal{M}_i})\backslash V]$, $V_{\mathcal{M}_i} = V_{\mathcal{M}_{i+}}$. The addition of some variable, X, to a given model, $\mathcal{M}_i$, is a 'mere inclusion' just in case adding X to $\mathcal{M}_i$ produces a new model, $\mathcal{M}_{i+}$, which is a conservative extension of $\mathcal{M}_i$. $\mathcal{M}_{i+}$ would be 'strictly richer' than $\mathcal{M}_i$.

This point does not always seem to be fully appreciated in the literature. For example, Hall says,

> Such model-relativity *might* make sense if what we are doing is choosing between different *ways* of representing a given aspect of some situation…. But it doesn't make sense if, instead, what we are doing in moving from one model to another is simply *increasing* the number of aspects we are choosing to represent. (2006, p. 34)

Similarly, Halpern writes,

> It seems that looking more and more carefully at a situation should not result in our view of $X = x$ being a cause of $Y = y$ alternating between "yes" and "no", at least, not if we do not discover anything inconsistent with our understanding of the relations between previously known variables. (2016b, p. 18)

But these comments miss the point that some cases of instability – such as that in $\mathcal{M}_3$, above – are perfectly acceptable. This local instability makes sense and is accommodated in the overall theory by a globally stable verdict.

Still, perhaps a different kind of instability would be worrying. This would be when a model delivers the local model-relative verdict that $X = 1$ actually causes $Z = 1$, but a strictly richer model delivers the verdict that $X = 1$ does not actually cause $Z = 1$. It would seem peculiar if merely zooming in on a causal relation could make it disappear. As we have seen, there is nothing wrong with a pattern where a series of strictly richer models deliver a 'no causation' verdict up until a certain point, after which models deliver a 'yes, causation' verdict. But it would indeed be concerning to have a pattern in which a model delivers a 'yes, causation' verdict but a strictly richer model delivers a 'no causation' verdict.

However, there is no reason to suppose that this kind of reversal can ever happen. We have looked at various cases in which it seems to – involving structural isomorphs and Halpern's series of alternating models – and we have seen how the requirement of Explicit Partial Mediation eliminates the troubling positive causal verdicts. In effect, by imposing this requirement we eliminate a range of model-interpretation pairs that would too easily generate 'yes, causation' verdicts. Now that these are off the table, I can think of no further cases in which a local positive causal verdict delivered by a model-interpretation pair will be reversed by an enrichment of that model.

3.4.c Stability Debunked

In sum, local instability of models *per se* is not a problem, since it is consistent with the stability of overall global verdicts. True, local instability in which enrichment reverses model-relative positive verdicts would be worrying, but there is no reason to suppose that this will ever occur once we impose the requirement of Explicit Partial Mediation. So I see no need to impose any extra Stability requirement on apt models. Explicit Partial Mediation alone does all the work that needs to be done.

CHAPTER 4

Modal Profiles and Causal Relativism

Abstract This chapter uncovers a heretofore hidden element in the interpretation of a causal model. It shows, first, that a given model on an interpretation can accurate or inaccurate of the same situation, depending on a further parameter. As I demonstrate, this further parameter is the set of background possibilities – what I call the *modal profile*. However, this observation raises a problem for a theory of actual causation in terms of these models. I conclude by proposing a view that takes this relativity at face value. According to it, actual causation holds only relative to a modal profile.

§4.1 Introduction

In the last chapter, we saw how accuracy is inadequate for aptness in permitting models that mistakenly affirm causation when there isn't any. These can be ruled out, however, by requiring Explicit Partial Mediation. So far, then, an apt model is one that is accurate and satisfies Explicit Partial Mediation. Yet this isn't the end of the story. As I will demonstrate, accuracy is not a determinate function of a model, an interpretation, and a situation. Since a given model on a given interpretation can still be deemed to be accurate or inaccurate of the same situation, there must be an additional parameter relative to which a model is accurate or not. I argue that this additional parameter is a specification of background possibilities, and I propose that such a specification be included in a model's interpretation.

So, a model represents a situation only relative to some modal profile or other. I conclude by exploring the ramifications this observation has for a theory of actual causation in terms of these models. I explain how, by existentially quantifying over models, the SEM definition existentially quantifies over all modal profiles. The problem is that some modal profiles deliver counterintuitive results. I look at three different ways to respond to this problem. Ultimately, I argue that the best response is to take this relativity at face value, treating actual causation as itself holding relative to a modal profile. This view has similarities and differences to Contrastivism about causation – the view that actual causation holds between a cause, an effect, and a set of contrast classes for each. But how deeply similar they really are depends on how details of each view are worked out.

§4.2 Accuracy as Relative

Consider again the definition of accuracy as laid out in Chapter 2. So far, an *interpretation* is defined as an assignment of a range of property instances to each variable of a model. An interpretation is *permissible* for representing some situation just in case it satisfies exclusivity, exhaustivity, and distinctness relative to that situation. A model will be accurate on an interpretation when the interpretation is permissible, the values assigned to the exogenous variables by the Assignment represent what really happened, and the counterfactuals entailed by the equations are true. Formally,

> **Accuracy:** A causal model, $\mathcal{M}_i$, on an interpretation $\mathcal{I}(\mathcal{M}_i)$, is accurate of a given situation, $\mathbb{S}$, just in case …
>
> i. $\mathcal{I}(\mathcal{M}_i)$ is a *permissible interpretation* of $\mathcal{M}_i$ for representing $\mathbb{S}$;
>
> ii. The content entailed by the assignment, $\boldsymbol{\mathcal{A}}_{\mathcal{M}_i}$, on $\mathcal{I}(\mathcal{M}_i)$ is the case in $\mathbb{S}$;
>
> iii. The counterfactuals entailed by the equations, $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$, on $\mathcal{I}(\mathcal{M}_i)$ are true in $\mathbb{S}$.

As I will show, whether (i) and (iii) are satisfied – that is, whether an interpretation is permissible and whether the entailed counterfactuals are true – depends not simply on the target situation, but on how the situation is *modally characterized* – how it is set against a background space of possibilities.

4.2.a The Relativity of the Permissibility of an Interpretation

Take first the permissibility of an interpretation. To illustrate the relativity of exclusivity and distinctness, consider:

> **Train and Two Tracks** There are two train tracks that separate at a certain point, then converge again before reaching the station. A lever at the separation point controls which track a train is sent down. A train travels down the left-hand track, arriving at the station.[1]

In **Train and Two Tracks**, a train can only travel down one of the two tracks. Since it can't be in two places at once, a train's traveling down the left-hand track entails that it does not travel down the right-hand track, and vice versa. So, a train being on the left-hand track is *not distinct* from its being on the right-hand track, and its being on one track *excludes* its being on the other. Relative to the space of possibilities in which there is only one train, the property-instantiation of the left-hand track being occupied is not distinct from that of the right-hand track being occupied. They are mutually exclusive alternatives. Thus, in order to satisfy exclusivity and distinctness, a permissible interpretation must represent these two property instances as two values of the same variable.

However, although there is in fact only one train, there could have been another. That a train travels down the left-hand track entails nothing about whether *another* train travels down the right-hand track, nor vice versa. Relative to a space of possibilities that permits the presence of a second train, the property instantiations of the left-hand track being occupied and of the right-hand track being occupied are *distinct*. A train's being on one track *does not exclude* another train's being on the other. Relative to this second space of possibilities, then, an interpretation that represents the left-hand track's being occupied and the right-hand

[1] This example can also be found in (J. Halpern & Hitchcock, 2010, pp. 398–399; Pearl, 2000, p. 324; Weslake, 2015, sec. 3.1; Woodward, 2016, pp. 1063–1064).

track's being occupied as two values of a single variable would violate both distinctness and exclusivity, and would therefore be impermissible. A permissible interpretation must represent these two property instances as values of different variables.

The principle of exhaustivity is also relative in this way to a space of possibilities. Consider:

> **Sophie in the Factory** Sophie the pigeon is trained to peck at and only at red things. Sophie lives in the yard of a paint chip factory that only produces scarlet and cyan paint chips. Sophie sees a scarlet paint chip in the yard and pecks at it.[2]

Relative to the space of possibilities constrained by being in the factory yard, the scarlet chip could only have otherwise been cyan. A binary variable that takes one value for scarlet and the other for cyan thus satisfies exhaustivity and is therefore *permissible* relative to this first space of possibilities. But relative to what is physically possible, the paint chip could have been any color. The binary variable, {*scarlet*, *cyan*}, fails to satisfy exhaustivity relative to this second space of possibilities, and is therefore *impermissible*.

Thus, whether an assignment of an underlying object and a range of properties to a variable is exclusive or exhaustive in a situation, or a set of such assignments distinct, depends on what possibility space is specified. This means that an interpretation is *permissible* only relative to possibility space, as is the accuracy of a model on an interpretation.

[2] This case is adapted from Yablo (1993).

4.2.b The Relativity of the Truth of the Equations

Indeed, the truth of the equations is also relative in this way. To illustrate, say we model **Sophie in the Factory** with the following model and interpretation:

$\mathcal{S}$ = **U** = {*X*}
V = {*Y*}
R = *f* (*X_i*) = {1, 0}

$\mathcal{A}$ = (EQ1) $X = 1$

$\mathcal{L}$ = (EQ2) $Y := X$

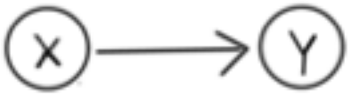


Figure 14. $\mathcal{M}_4$

$\mathcal{I}(\mathcal{M}_4)_{SF}$: *X (paint chip) :=* $\begin{cases} 1 \; if \; scarlet \\ 0 \; if \; not \; scarlet \end{cases}$

Y (Sophie's action) := $\begin{cases} 1 \; if \; pecks \\ 0 \; if \; doesn't \; peck \end{cases}$

Is $\mathcal{M}_4$ accurate of **Sophie in the Factory** on $\mathcal{I}(\mathcal{M}_4)_{SF}$? First, (1) is satisfied. The chip's being scarlet and its being not scarlet are exclusive as well as exhaustive alternatives, Sophie pecking and not pecking are exclusive as well as exhaustive alternatives, and the chip's being scarlet or not is relevantly distinct from Sophie's pecking or not. Second, (2) is satisfied. The Assignment sets *X* to 1, which represents the paint chip being scarlet, which it is in **Sophie in the Factory**. Finally, is (3) satisfied? The counterfactuals entailed by $\mathcal{M}_4$ on $\mathcal{I}(\mathcal{M}_4)_{SF}$ are:

(i) If the chip were scarlet, then Sophie would peck.

(ii) If the chip were not scarlet, then Sophie would not peck.

First, (i) is true. Intervening on the situation to set the chip to scarlet would result in Sophie pecking. Is (ii) true? Surprisingly, it depends. If we hold fixed the way this factory really operates, then the only way a chip could fail to be scarlet in this factory yard is if it were cyan. And if it were cyan, then Sophie would not peck. So, when we allow what's possible to be constrained by contingent background facts, (ii) comes out true. Therefore, $\mathcal{M}_4$ is accurate of **Sophie in the Factory** on $\mathcal{I}(\mathcal{M}_4)_{SF}$ *on the space of possibilities constrained by how the factory actually operates*.

But it is not accurate tout court. If we allow that the paint chip could have been any physically possible color, (ii) is false. Some permissible interventions on the situation will set the chip to a non-red color, in which case Sophie would not peck. But many permissible interventions will set the chip to a non-scarlet shade of red. And if the chip had been any non-scarlet red color, then Sophie would still have pecked. If we assume a kind of universal principle whereby *every* permissible intervention must be such that the consequent comes out true, then (i) is false. And such a universal principle seems to be the standard assumption on the rare occasion this issue of multiple possible interventions is addressed in the literature.[3] Thus, $\mathcal{M}_4$ is *not* accurate of **Sophie in the Factory** on $\mathcal{I}(\mathcal{M}_4)_{SF}$ *on the space of possibilities constrained by physical possibility*.

[3] See Briggs (2012, 152 -3).

Thus, the counterfactuals entailed by a model on an interpretation may be true relative to one space of possibilities, yet false relative to another. This is because which domain of possible worlds is specified may change *how* a given property can be instantiated by some particular object, and so change whether some counterfactual involving the instantiation of that property is true. An alternative way to put this point is that the truth of the equations is relative because they do not yet entail counterfactuals when interpreted by an assignment of ranges of property instances to the variables. Counterfactuals are entailed only once a space of possibilities is also specified. Whether an equation is true or false may then change depending on which counterfactual it really entails. On this view, the entailed counterfactual's *truth-conditions* are relative to a space of possibilities, rather than their *truth-values*.[4]

§4.3 Modal Profiles

So, accuracy of a model on an interpretation is relative to a space of possibilities. **Train and Two Tracks** and **Sophie in the Factory** can be modelled relative to one or the other, changing whether a given interpretation is permissible and changing the truth values of the entailed counterfactuals. It is a further question whether some space of possibilities is such that a situation *should* be represented relative to it and not the other. This issue is addressed in §4.5 and §4.6.

[4] Thanks to Jonathan Schaffer for this alternative way of putting the point.

4.3.a Coining/ Revising "Modal Profile"

For ease of exposition, I call a space of possibilities a *modal profile*. This is a slight revision on the term "modal profile" as it is standardly used. I take the following quote to be illustrative: "A[n object's] modal profile (or nature) captures all the possible combinations of properties the object might instantiate in different possible worlds. (Schroeter, 2019, n. 2)" In this sense, the "modal profile" of a situation is the complete story of how things in that situation could have been or gone – all the possible combinations of properties possibly instantiated by the objects in that situation. But notice that holding certain features of the situation fixed will rule out those combinations of properties with which these fixed features are incompatible, thus restricting the possibility space. When we hold fixed the background fact of there being only one train, this rules out the possibility of both tracks being occupied at the same time. Holding different features fixed will rule out different combinations of properties, thus restricting the possibility space in different ways. This means that for any situation there is a whole family of possible restrictions on how that situation could have gone, each member of which results from holding fixed a different set of facts about that situation.

It has long been appreciated that such restrictions play a crucial role in evaluating counterfactuals. Which restriction is the right one for evaluating a counterfactual is arguably the fundamental question that a semantics of counterfactuals needs to answer.[5] I am here demonstrating that such restrictions play a crucial role in how a causal model represents, as

[5] I discuss this question further in Chapter 6, specifically §6.2.c.

well, regardless of whether they are taken to represent counterfactual dependencies or type-level causal dependencies. Due to their significance, it will be helpful to permit the term *modal profile* to refer to any possible combination of properties possibly instantiated by the objects in the situation – even if the combination of properties results from holding certain features fixed. We can still use the expression *universal modal profile* to refer to *all* possible combinations. In my sense, then, situations or objects do not have a single modal profile but a family of them.

So, a modal profile, as I use the term, picks out a portion of modal space – i.e. specifies a domain of possible worlds. Thus, the idea of relativizing a model to a modal profile means that a variable partitions a *portion* of modal space – the portion picked out by the modal profile – mapping its values one-to-one onto each unit of the partition.

4.3.b Existing Sensitivity to Modal Profiles: Relativity and Serious Possibilities

While the full extent of the relativity of models to modal profile has yet to be recognized in the literature on causal models, there is a degree of sensitivity to it. For example, in their seminal text on structural equation models, Sprites, Glymour, and Scheines explain that type-level causal relations are relative to what is taken to be possible as part of the background. They write,

> If our notion of causation between variables were strictly applied, almost every natural variable would count as a [type-level] cause of almost every other natural

> variable, for no matter how remote two variables, A and B, may be, there is usually *some* physically possible – even if very unlikely – arrangement of systems such that variation in some values of A produces variation in some values of B….In practice, we always consider a restricted range of variation of other variables in judging whether A causes B. Strictly, therefore, our definitions of causal relations for variables should be relative to a set of possible values for other variables, but we will ignore this formality and trust to context. (1993, p. 44)

Sprites, Glymour, and Scheines here recognize the implicit relativity of models to background possibilities when the models are used to capture type-level causation. Notice that this relativity carries over to the use of these models to define actual causation, as well – at least for those who interpret the equations of a model to represent type-level causal dependencies.

A sensitivity to this relativity can also be seen in discussions of "serious" exhaustivity, with which I engaged briefly in Chapter 2 (§2.3). This version of exhaustivity requires that values of variables represent all *and only* those possibilities that we are willing to take seriously. This principle serves the function of ruling out certain modal profiles – namely, those that include non-serious possibilities in the ranges of alternatives represented by variables. It can therefore be taken as indicative of a sensitivity to the fact that causal relations are represented in a model as holding relative to a background space of possibilities, and that certain spaces of possibilities might not be of the right kind.

I take my idea of relativity to modal profiles to be an improvement on this qualified version of exhaustivity for two reasons. I will argue fully later, in §4.6.d, for the first reason. For now, I will merely claim that relativity to modal profiles can not only do the work that serious exhaustivity is meant to do, but it can do this work better – in a way that is less threatening to realism about actual causation. The second reason is more obvious. This is that relativity to modal profiles can capture not only the relativity of *exhaustivity* to a background space of possibilities, but that of exclusivity and distinctness, as well. It is therefore a more general principle, from which the "serious" qualification on exhaustivity may follow.

4.3.c Including Modal Profiles in the Interpretation

As demonstrated, whether a model is accurate depends in part on which of many possible modal profiles is presupposed. But there is nothing yet in the model or interpretation to indicate which one this is. The modal profile is therefore, strangely enough, an as yet unrecognized, *additional* element of how causal models represent. The next question, then, is how to incorporate it so as to render determinate our definition of accuracy. Two ways suggest themselves.

The first way would be to enrich what is meant by "situation". I have thus far been using "situation" in a pre-theoretic sense to pick out some real-world situation – roughly, to denote a chunk of spacetime.[6] But we could build the modal profile into the delineation of a situation,

[6] I leave open questions to do with what comprises a chunk, such as whether a chunk is necessarily coherent. How these questions should be answered will depend on the purposes to which we want to put situations. Thus, a full discussion will take me too far afield.

explicating “situation” so as to include an indication of which features are presumed fixed, which features are permitted to vary, and in what ways. As a result, $\mathbb{S}$ would already bring with it a particular modal profile relative to which the variables of an accurate model would need to satisfy exclusivity, exhaustivity, and distinctness, and relative to which its entailed counterfactuals would need to be true.

The second way to incorporate the modal profile would be to enrich the interpretation. This way keeps the pre-theoretic understanding of “situation” as referring only to a particular chunk of spacetime. As a result, the very same situation can be modelled relative to different modal profiles. Which one is presupposed by a model is specified by an extra component of the interpretation. Thus, an interpretation becomes an assignment of content *and a specification of modal profile*. I will adopt this second understanding, in part because it is a more natural use of the term “situation”, but more importantly because it is a more perspicuous one. It permits us to talk easily and clearly about both raw, undifferentiated situations and the different corresponding modal profiles. A modal profile can be exhaustively specified by defining a set of possible worlds, or it can be quickly specified by enumerating the features of a situation which are held fixed (implying that the unenumerated features are permitted to vary in accord with what is metaphysically possible given the fixed features). I will normally do the latter.

This calls first for an update to what constitutes a permissible interpretation. $\mathcal{I}(\mathcal{M}_i)$ is a *permissible interpretation* of $\mathcal{M}_i$ for representing $\mathbb{S}$ just in case the content entailed by the signature, $\boldsymbol{\mathcal{S}}_{\mathcal{M}_i}$, given $\mathcal{I}(\mathcal{M}_i)$ satisfies exclusivity, exhaustivity, and distinctness relative to the modal profile specified by $\mathcal{J}(\mathcal{M}_i)$. In addition, this calls for an amendment to the accuracy condition pertaining to the equations. The new condition reads: the counterfactuals entailed by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$ on $\mathcal{J}(\mathcal{M}_i)$ are true in $\mathbb{S}$ *relative to the modal profile specified by* $\mathcal{J}(\mathcal{M}_i)$.

Putting this all together results in our final definition of accuracy:

> **Accuracy:** A causal model, $\mathcal{M}_i$, is accurate of a given situation, $\mathbb{S}$, on an interpretation $\mathcal{J}(\mathcal{M}_i)$, just in case …
>
> i. $\mathcal{J}(\mathcal{M}_i)$ is a permissible interpretation of $\mathcal{M}_i$ for representing $\mathbb{S}$;
>
> ii. The content entailed by $\boldsymbol{\mathcal{A}}_{\mathcal{M}_i}$ on $\mathcal{J}(\mathcal{M}_i)$ is the case in $\mathbb{S}$;
>
> iii. The counterfactuals entailed by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$ on $\mathcal{J}(\mathcal{M}_i)$ are true in $\mathbb{S}$ relative to the modal profile specified by $\mathcal{J}(\mathcal{M}_i)$.

4.3.d Merely an Illusion of Relativity?

One might be tempted to object that this relativity is merely an illusion that results from our use of simplified models which, in turn, is merely due to our calculative limitations. The relativity would dissolve, so this objection goes, if we were able to use a *universal* model. Such a model would represent the single, objective story about all the possible ways things could go. It would therefore be accurate of any given situation tout court, dissolving any relativity to modal profiles and obviating the need to incorporate this relativity into our theory of aptness (except perhaps for pragmatic reasons).

There are two problems with this suggestion. The first is that, even if a universal model were in principle possible, it would deliver the verdict that almost every actual feature of reality is an actual cause of almost every other actual feature.[7] To reiterate a portion of the earlier quote from Spirtes, Glymour, and Scheines,

> If our notion of causation between variables were strictly applied, *almost every natural variable would count as a [type-level] cause of almost every other natural variable*, for no matter how remote two variables, A and B, may be, there is usually *some* physically possible – even if very unlikely – arrangement of systems such that variation in some values of A produces variation in some values of B. (1993, p. 44, emphasis my own)

While Spirtes, Glymour, and Scheines are focused on type-level causation, this point applies to actual causation, as well. Take almost any two property instantiations, F_1a and G_1b, however intuitively unrelated, where G_1b lies in F_1a's forward light cone. There will be an alternative *albeit perhaps highly unlikely* property, F_2, that *a* could have instantiated such that, had it been the case that F_2a, then *b* would have instantiated an alternative property, G_2. For example, take a particular car being parked on May 21, 1932 in the suburbs of Seattle, WA and the COVID-19 pandemic beginning in December 2019 in Wuhan, China. Had that car instead instantiated the property of time-travelling through a wormhole to Wuhan and exploding in the right spot and right time to kill patient zero before the virus spread, then the pandemic would not have happened. While a wildly remote possibility, the instantiation

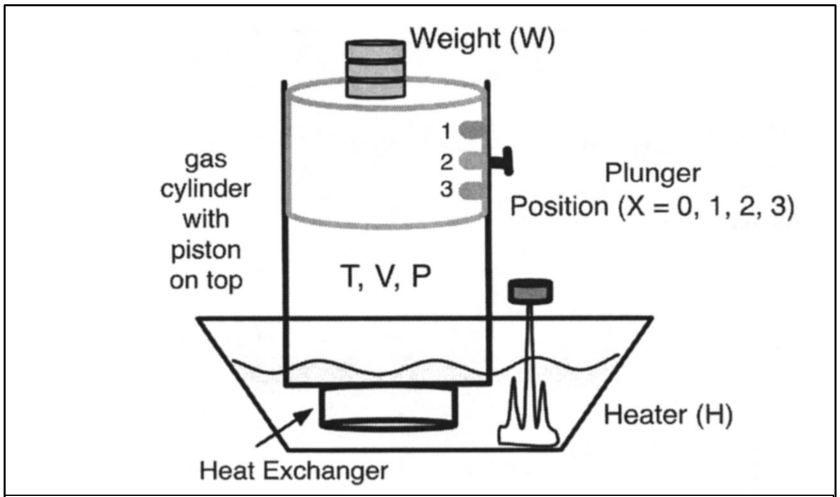
[7] Thanks to Jonathan Schaffer for suggesting this response.

of such a property is nevertheless consistent with our best physical theories. If all possibilities are to be included, then this possibility should be included. The result would be that this car being parked on this day in this spot is a *direct* actual cause of the COVID-19 pandemic. Indeed, *every* property instantiation in the backwards light cone of the dawn of COVID-19 is a direct actual cause of the pandemic. In general, a universal model would make it so that every property instance in F_1a's backwards light cone is a direct actual cause of F_1a, *and* make it so that F_1a is a direct actual cause of *every* property instance in its forward light cone.[8]

The second problem with this suggestion is that, as it turns out, a universal SEM or DAG is not possible. The best argument for this comes from Hausman, Stern, and Weinberger (2014). They point out that by altering the initial conditions of some physical systems, the causal structure of the system may be altered in a way that cannot be captured by a SEM or DAG. They focus on the following case, which they call **The Elusive Cylinder**. Consider:

> [The device in Fig. 15] consists of gas immersed in a water bath that is maintained at a constant temperature H. There is a piston at the top of the cylinder that can be locked into one of three positions ($X = 1, 2,$ or, 3) or allowed to move up or down depending on the pressure of the gas ($X = 0$) and on the weight placed on top of the piston. (Hausman et al., 2014, p. 1926)

[8] An even worse problem would result were we to take seriously *backwards* time travel, as well. Whether this is consistent with our best physical theories remains a point of controversy, however.


Weight (W)
1
2
3
gas cylinder with piston on top
Plunger Position (X = 0, 1, 2, 3)
T, V, P
Heater (H)
Heat Exchanger

The Elusive Cylinder (Hausman et al., 2014, p. 1926)

Figure 15.

The key feature of this case is that the dependency relations that hold between the property instantiations in this system are *structurally* different when the piston is locked (*X ≠ 0*) than when it is not locked (*X = 0*). This means that when we try to represent an intervention on the system that changes the piston from locked to unlocked, or back again, the result of the intervention will not simply propagate throughout a single model's system of equations, changing the values of variables in line with the equations. This is how the results of interventions are normally represented. Instead, an intervention on the piston will result in a *change in the equations themselves*. To see this, compare the two sets of equations and corresponding DAGs that result from *X ≠ 0* and from *X = 0*, respectively. For both, the signature is the same: the set of exogenous variables, ***U*** = {*H, W, X*} and the set of endogenous variables, ***V*** = {*P, T, V*}.

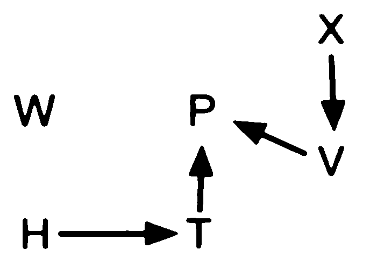
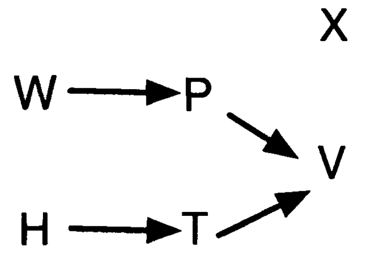
	SEM Equations	DAG
X ≠ 0	$\mathcal{L}_{\mathbf{X\neq0}}$ = (EQ1) $T := f_T(H)$ (EQ2) $P := f_P(T, V)$ (EQ3) $V := f_V(X)$	
X = 0	$\mathcal{L}_{\mathbf{X=0}}$ = (EQ1) $T := f_T(H)$ (EQ2) $P := f_P(W)$ (EQ3) $V := f_V(P, T)$	

Figure 16.

Say that the piston is not locked. The correct assignment of a representative model would set $X = 0$, and the resulting set of equations would be $\mathcal{L}_{\mathbf{X=0}}$. But within such a model, the "variable" *X* that takes the value *0 cannot* take a different value. The inclusion of a particular value for a variable is standardly taken to imply the possibility of intervening on that variable to set it at that value. But the result of such an intervention on *X* to change its value from $X = 0$ to $X \neq 0$ would be a change in the equations! The formalism cannot accommodate this. Therefore, the "variable" *X* can only take one value – *0*. But this means that it is not a proper variable, since a variable must be able to take at least *two* values. The piston being unlocked must be represented as a background condition – as part of the modal profile – implied by a selection of the equations in $\mathcal{L}_{\mathbf{X=0}}$.

So, it will not be possible for a model that represents what would happen given an unlocked piston to simultaneously represent what would happen were the piston to be locked – were $X \neq 0$. However, for a mechanical system of this type with an unlocked piston, there is still the physical *possibility* that the piston be locked. The crucial point here is that this possibility cannot be represented in the same model that represents the actual conditions of the system. No single model can say what would happen, given an unlocked piston, *and* what would happen were the piston to be locked.

I take this example to carry a general moral. Some possible changes in features of the world change the relationships between other features. In model terms, this means that some interventions on variables change the equations. But this cannot be captured in the SEM formalism. Thus, even if there is a single, objective story about all the possible ways things could go, there can be no single, universal model that captures it.

§4.4 The Problem of Counterintuitive Verdicts

The need to incorporate modal profiles into how a causal model represents has widespread theoretical ramifications. Most interesting is a problem it raises for extant theories of actual causation in terms of these models. So far, no constraints have been placed on which modal profiles might be eligible to figure in an apt model-interpretation pair. However, some modal profiles figure in otherwise apt model-interpretation pairs that nevertheless produce counterintuitive causal verdicts. This raises a problem for any SEM definition of actual causation. For simplicity, though, I will focus on the problem it raises for those definitions, such as $\boldsymbol{AC - simpliciter}$, that *existentially* quantify over apt model-interpretation pairs, and so existentially quantify over any modal profile that figures in such a pair. There are three kinds of cases that deliver especially counterintuitive causal verdicts. I will focus on these, providing an example of each.

4.4.a Overly General Causes

In the first kind of case, overly general things qualify as causes. For example, take a situation similar to what we had before with **Sophie in the Factory**, but instead of Sophie, consider Alice the pigeon.

> **Alice in the Factory** Alice the pigeon is trained to peck at and only at *scarlet* things. Alice, like Sophie, lives in the yard of a paint chip factory that only produces scarlet and cyan paint chips. Alice sees a scarlet paint chip in the yard and pecks at it.[9]

Say we model **Alice in the Factory** with $\mathcal{M}_4$ (reprinted below) on $\mathcal{I}(\mathcal{M}_4)_{AF}$.

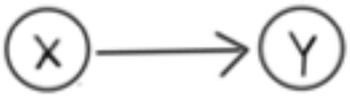
[9] This is adapted from a case due to Shoemaker (2003), which in turn is adapted from the case referenced above, in Yablo (1993).

Figure 14. (again)

$\boldsymbol{\mathcal{S}}$ = **U** = {X}
V = {Y}
R = $f(X_i)$ = {1, 0}

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $Y := X$



$\boldsymbol{\mathcal{M}_4}$

$\mathcal{I}(\mathcal{M}_4)_{AF'}$: *X (paint chip) :=* $\begin{cases} 1\ if\ red \\ 0\ if\ not\ red \end{cases}$

Y (Alice's action) := $\begin{cases} 1\ if\ pecks \\ 0\ if\ doesn't\ peck \end{cases}$

Modal Profile: holds fixed being in the factory yard

First, I will quickly confirm that $\mathcal{M}_4$ on $\mathcal{I}(\mathcal{M}_4)_{AF}$ is accurate of **Alice in the Factory**. The content assigned to the variables satisfies exclusivity, exhaustivity, and distinctness and the assignment says truly that the chip was red. Finally, the entailed counterfactuals are true. Had the chip not been red, then Alice would not have pecked and, given the modal profile, had the chip been red, then Alice would have pecked – since the only way the chip could be red is by being scarlet.

Next, notice that $\mathcal{M}_4$ says that *X = 1* is an actual cause of *Y = 1*, and $\mathcal{I}(\mathcal{M}_4)_{AF'}$ interprets this to mean that the chip's being *red* is an actual cause of Alice pecking. And since there is at least one apt model-interpretation pair that delivers this verdict, $\boldsymbol{AC - simpliciter}$ says that the chip's being red *just is* an actual cause of Alice pecking. It is an actual cause *simpliciter*.

However, this result seems counterintuitive. Red is too general to be a cause. The chip instantiating the property of being red is only causally efficacious because it happens to be coextensive with the property instantiation of being scarlet, given that the fact of its being in the factory yard is held fixed. Intuitively, though, it could have been red without being scarlet and, had that been the case, then Alice would *not* have pecked. This result is at minimum highly misleading. [10]

4.4.b Irrelevant *Positive* Causes

The second and third kinds qualify prima facie irrelevant things as causes: either prima facie irrelevant *positive* property instantiations or prima face irrelevant *omissive* property instantiations.[11] As an example of the former, consider the following situation. While it is fairly involved, it will serve as an example of the latter, as well, so bear with me.

[10] I should point out that there is, in fact, a fourth kind of case that delivers counterintuitive causal verdicts – where overly *specific* things qualify as causes. But the verdicts delivered in these cases strike me as far less counterintuitive than those delivered in the others. As a quick example, say we model **Sophie in the Factory** in the same way as above, in §4.2.b, with $\mathcal{M}_4$ on $\mathcal{I}(\mathcal{M}_4)_{SF}$. Add to $\mathcal{I}(\mathcal{M}_4)_{SF}$ a specification of modal profile that holds fixed Sophie's being in the factory yard. This model-interpretation pair is accurate of **Sophie in the Factory** and, according to it, the chip's being *scarlet* caused Sophie to peck. $\boldsymbol{AC - simpliciter}$ therefore says that the chip's being scarlet *just is* a cause of Sophie pecking. However, the chip's being scarlet is too specific. Intuitively, the chip could have been a non-scarlet shade of red. If it were, then it would not be true that had the chip not been scarlet, then Sophie would not have pecked. This result is misleading. But, as I said, it strikes me as less misleading than the others.

[11] This pair of problems comes from (Sartorio, 2010), who presents them as a problem for counterfactual accounts of causation generally. She calls them the *Problem of Unwanted Positive Causes* and the *Problem of Unwanted Negative Causes*. The case of **The Prince and his Biscuits** is adapted from an example she provides.

> **The Prince and his Biscuits** The Queen of England has to be out for the day. She asks the Prince of Wales to water her plant in her absence, which is fragile and needs to be watered daily in order to survive. The Prince promises to water it but eats biscuits instead. The plant dies.

Suppose the following also holds of **The Prince and his Biscuits**: the Prince has been raised to be a man of his word, who always acts on a promise. However, he is uniquely akratic when it comes to biscuits. He had forgotten when making his promise that today was Biscuit Day – the one day a year when biscuits are put out in the tearoom on the far side of the palace from the greenhouse. Furthermore, there is an automatic locking mechanism on the greenhouse that unlocks the room only from 12:02 to 12:04, for the plant's protection, and yet the biscuits are put out during this very same span of time! Finally, it would take even the fastest runner 10 minutes to get from the greenhouse to the tearoom, or back again.

This situation can be accurately modelled with $\mathcal{M}_4$, from earlier, on the following interpretation:

$\mathcal{I}(\mathcal{M}_4)_{PB}$: *X (Prince of Wales) :=* $\begin{cases} 1\ if\ eats\ biscuits \\ 0\ if\ waters\ plant \end{cases}$

Y (plant) := $\begin{cases} 1\ if\ dies \\ 0\ if\ survives \end{cases}$

Modal Profile: holds fixed Prince's character, lock mechanism,
holiday schedule, and palace layout

First, $\mathcal{M}_4$ on $\mathcal{I}(\mathcal{M}_4)_{PB}$ is apt for representing **The Prince and his Biscuits**. $\mathcal{I}(\mathcal{M}_4)_{PB}$ is permissible relative to the specified modal profile, the assignment says truly that the Prince ate biscuits, and the counterfactuals entailed by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_4}$ are true. Then, according to $\mathcal{M}_4$, *X = 1* actually causes *Y = 1*, and $\mathcal{I}(\mathcal{M}_4)_{PB}$ interprets this to mean that the Prince's eating biscuits is an actual cause of the plant dying. And since there is at least one apt model-interpretation pair that delivers this verdict, $\boldsymbol{AC - simpliciter}$ says that the Prince's eating biscuits *just is* an actual cause of the plant dying. It is an actual cause *simpliciter*.

But this also seems counterintuitive. The Prince's eating biscuits is prima facie irrelevant to the dying of the plant. The Prince instantiating the property of eating biscuits is only causally efficacious because it happens to be mutually exclusive with his watering the plant, given that his character, the locking mechanism, the holiday schedule, and the layout of the palace are held fixed. Intuitively, though, he could have watered the plant while eating biscuits and, had that been the case, then the plant would not have died. Further, there's an intuitive sense in which he could have not eaten the biscuits but also failed to water the plant. Had that been the case, then the plant *still* would have died. This result is also at minimum highly misleading.

4.4.c Irrelevant *Omissive* Causes

We can use this same example to illustrate the final kind of case: how modal profiles qualify prima facie irrelevant *omissive* property instantiations as causes. Imagine that the biscuits eaten by the Prince then give him a stomach ache. Call this enhanced case **The Prince's Stomach Ache**. We can model this again with $\mathcal{M}_4$, on the following new interpretation:

$$\mathcal{I}(\mathcal{M}_4)_{PSA}: \quad \textit{X (Prince of Wales) :=} \begin{cases} 1\ if\ doesn't\ water\ plant \\ 0\ if\ waters\ plant \end{cases}$$

$$\textit{Y (Prince's stomach) :=} \begin{cases} 1\ if\ develops\ ache \\ 0\ if\ doesn't\ develop\ ache \end{cases}$$

Modal Profile: holds fixed Prince's character, lock mechanism, holiday schedule, and palace layout

First, $\mathcal{M}_4$ on $\mathcal{I}(\mathcal{M}_4)_{PSA}$ is apt for representing **The Prince's Stomach Ache**. $\mathcal{I}(\mathcal{M}_4)_{PSA}$ is permissible relative to the specified modal profile, the assignment says truly that the Prince didn't water the plant, and the counterfactuals entailed by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_4}$ are true. Then, according to $\mathcal{M}_4$, *X = 1* actually causes *Y = 1*, and $\mathcal{I}(\mathcal{M}_4)_{PSA}$ interprets this to mean that the Prince's not watering the plant is an actual cause of him developing a stomach ache. And since there is at least one apt model-interpretation pair that delivers this verdict, $\boldsymbol{AC-simpliciter}$ says that the Prince's not watering the plant *just is* an actual cause of his developing a stomach ache. It is an actual cause *simpliciter*.

This, too, seems counterintuitive. The Prince's not watering the plant is prima facie irrelevant to his developing a stomach ache. The Prince instantiating the property of not watering the plant is only causally efficacious because it is coextensive with his eating biscuits, given that his character, the locking mechanism, the holiday schedule, and the layout of the palace are held fixed. Intuitively, though, he could have failed to water the plant while also failing to eat the biscuits and, had that been the case, he would not have developed a stomach ache. Furthermore, there's an intuitive sense in which he could have watered the plant *and* eaten the biscuits and, had that been the case, then he still would have developed a stomach ache. Again, this result is at minimum highly misleading.

§4.5 Two Initial Responses

The problem, then, is that some modal profiles qualify as causes things that don't seem to intuitively strike us as causes. I will first introduce and dismiss two initial responses one might make, before proposing and exploring the response that I endorse and the view of actual causation that results.

4.5.a Providing an Error Theory

A first possible response is to bite the bullet on any counterintuitive verdicts delivered. On the view that follows, it just is the case that the chip's being red is an actual cause simpliciter of Alice pecking, that the Prince's eating biscuits is an actual cause simpliciter of the plant dying, and that the Prince's not watering the plant is an actual cause simpliciter of his developing a stomach ache. How compelling this view is will depend on the strength of the error theory it can provide. As a first pass, we might say that these causal verdicts strike us as counterintuitive because the modal profiles responsible for making them the case are rare and infrequently encountered. The argument here would be that the cognitive apparatus that gives rise to causal judgment evolved so as to produce judgments relevant for the domain of modal profiles that we are likely to encounter. It is not designed to deal with unusual modal profiles and therefore misfires when faced with them.

One advantage of this view is that it preserves both realism about and objectivity of causation. Realism is preserved because there has been no invocation of mind or language dependent considerations in defining actual causation. The SEM definition simply quantifies over all apt model-interpretation pairs, and aptness has been defined without recourse to anything pragmatic. The sense of 'objective' that I have in mind here is the property ascribed to an area of study when questions about that area have determinate answers.[12] Actual causation is an objective area in this sense insofar as the answer to whether a particular thing, *c*, is an actual cause of particular thing, *e*, is an absolute 'yes' or 'no'. That causation is objective in this sense follows from any view that does away with reference to models, variable sets, modal profiles, etc. Objectivity about causation follows from this first view since it quantifies over all apt model-interpretation pairs to deliver verdicts about causation simpliciter. It preserves realism and objectivity, however, at the cost of sanctioning pervasive disagreement between actual causation verdicts delivered by our theory and those delivered by intuition.

4.5.b Supplementing Aptness

[12] While normally run together, realism and objectivity are substantively different. See (Clarke-Doane, 2020, p. 27) for this way of thinking about realism and objectivity.

A second initial response is to try to rule out in some way those modal profiles relative to which counterintuitive causal verdicts come out. The obvious way to go about this would be to supplement the theory of aptness, which would then serve to whittle down the domain of modal profiles over which the SEM definition quantifies. The resulting view would preserve the objectivity of causation – by quantifying over all apt model-interpretation pairs in line with the first view – and would seem to comport with intuition.

Unfortunately, this view seems to preclude realism about causation. I am inclined to think that the only satisfactory way to rule out all offending modal profiles would have to call upon pragmatic considerations. But then the resulting theory is not a realist one. Some in the literature seem happy to relinquish realism about actual causation (N. Hall, 2007; J. Halpern & Hitchcock, 2010). I think this is too great a concession and, in fact, too quick. I am open to the possibility that there exists an as yet undiscovered, mind and language independent supplement to aptness that successfully rules out offending modal profiles, though I would want to see the details. However, there is one final, alternative response to explore.

§4.6 Causal Relativism: Relativizing Actual Causation to Causal Profile

So, the first view accepts the counterintuitive verdicts and provides an error theory, preserving realism about actual causation at the cost of widespread tension between the results of the theory and causal intuition. The second view blocks the counterintuitive verdicts by restricting the domain of eligible modal profiles, but doesn't seem able to do so without giving up on realism. This takes me to my preferred response – to incorporate relativity to modal profile directly into the metaphysics of actual causation. I call this *Causal Relativism* about actual causation. We can motivate this view by showing how it makes sense of our intuitive causal judgments better than either of the previous two views while still preserving realism about actual causation.

4.6.a Fuzzy Intuitions Made Clear

Consider more carefully the counterintuitive verdicts previously laid out – for example, that the chip's being red is an actual cause simpliciter of Alice pecking. This does seem wrong. But that does not necessarily mean that it is wrong *simpliciter* that the chip's being red is an actual cause of Alice pecking. It's true that there is a sense in which the chip's being red is not an actual cause of Alice pecking. But there is also a sense in which it *is*. It makes sense to say that the chip's being red is *not* an actual cause of Alice pecking *given the fact that the chip could have been red without being scarlet.* But it also makes sense to say that the chip's being red *is* an actual cause of Alice pecking *given the fact that any red chip in the factory yard is a scarlet chip*.

It strikes me that the real problem with either of the previous two views, and with any view that quantifies over apt model-interpretation pairs, is that they leave out a crucial part of the story – namely, what background possibilities are in place. In short, these causal verdicts are not straightforwardly counterintuitive. Intuition about these cases instead feels *fuzzy*, and the fuzziness is resolved when we make explicit the background possibilities.[13]

[13] Thanks to Michael Strevens for this way of putting things.

Applying this to **The Prince and his Biscuits**, it seems wrong to say that the Prince's eating biscuits is an actual cause *simpliciter* of the plant dying. But it makes sense to say that the Prince's eating biscuits *is* an actual cause of the plant dying *given the Prince's character, the lock mechanism, the holiday schedule, and the layout of the palace*, and that the Prince's eating biscuits *is not* an actual cause of the plant dying *given the fact that it is metaphysically possible for him to both eat biscuits and water the plant.*

Finally, as applied to **The Prince's Stomach Ache**, it seems wrong to say that the Prince's not watering the plant is an actual cause *simpliciter* of his developing a stomach ache. But it makes sense to say that his not watering the plant *is* an actual cause of his developing a stomach ache *given the Prince's character, the lock mechanism, the holiday schedule, and the layout of the palace*. In addition, it makes sense to say that the Prince's not watering the plant *is not* an actual cause of his developing a stomach ache *given the fact that it is metaphysically possible for him to both fail to water the plant and fail to eat biscuits.*

4.6.b The Proposal: Causal Relativism

These intuitions can be taken at face value by treating actual causation as itself holding relative to a modal profile. While there are different ways one might translate this into a broader theory, I propose that the relativity of actual causation to modal profile be treated on analogy with the relativity of simultaneity to reference frame. It follows from the theory of special relativity that one event is not simultaneous with a second event *simpliciter*, but only *relative to an inertial reference frame* – an assignment of coordinates to events.

Simultaneity is therefore not a two-place relation but a *three-place* one, holding between one event, a second event, and an inertial reference frame. Analogously, I propose that actual causation is not a two-place relation that holds between a particular cause and effect, but a three-place relation that holds between a particular cause, an effect, and a modal profile. I call this view *Causal Relativism*. Incorporating relativity to modal profile into our SEM definition of actual causation gives us the following:

> $(\boldsymbol{AC - modal})$ *c* is an actual cause of *e* relative to a modal profile, θ_i, just in case there is an apt model-interpretation pair, <$\mathcal{M}_i$, $\mathcal{I}(\mathcal{M}_i)$>, where $\mathcal{I}(\mathcal{M}_i)$ specifies θ_i and represents *c* as *X* = *x* and *e* as *Y* = *y*, and $\mathcal{M}_i$ delivers the ***AC – relative*** verdict that *X* = *x* is an actual cause of *Y* = *y*.

According to Causal Relativism, causation is not objective in the sense that there is *no uniquely correct* causal structure. There are many correct structures, and F_ia may be an actual cause of G_ib relative to one structure but *not* an actual cause relative to a second. The view is therefore a kind of causal relativism. There is no absolute fact of the matter as to what actually causes what *simpliciter*. It is a relative matter. Facts about what actually causes what are relative to a modal profile. This is a disadvantage to the theory insofar as we find compelling the claim that causation is a determinate matter full stop. But it strikes me as no great loss given determinacy is recovered once the modal profile is filled in.

We can explain the intuitive fuzziness of the verdicts delivered by our theory as resulting from an ambiguity between two ways of filling in the hidden parameter of modal profile. The fuzziness is resolved once the modal profile is made explicit. The causal claims "the chip's being red is an actual cause of Alice pecking," "the Prince's eating biscuits is an actual cause of the plant dying," and "the Prince's not watering the plant is an actual cause of his developing a stomach ache" are incomplete. It is only once a modal profile is filled in – either implicitly or explicitly – that a causal claim has determinate truth-conditions. In general, claims of the form '*c* is an actual cause of *e*' have determinate truth-conditions only when the hidden parameter of modal profile is filled in.

4.6.c Contrastivism

This view has an affinity to *Contrastivism* – the view about actual causation whereby the causal relata are taken to be contrastive in nature [14] – although I will claim that while Causal Relativism can do the work done by Contrastivism, the converse does not hold. The general idea behind Contrastivism is that, despite superficial appearances, actual causation is *not* a binary relation that holds between a token cause and token effect. Instead, it is a four-place relation that holds between a cause, an effect, and a contrast class for each.[15] So, rather than actual causation as a relation of the form '*c* causes *e*', it is taken to have the form: '*c* rather than *c** caused *e* rather than *e**,' where the contrasts referred to by '*c** ' and '*e** ' may be singleton sets, but may also be many-membered sets. To briefly motivate this view, consider the following causal claim:

[14] See, for example, (Hitchcock, 1996b, 1996a, 2011; Maslen, 2004; Northcott, 2008; Reiss, 2013a, 2013b; Schaffer, 2005, 2010, 2012; Sinnott-Armstrong, 2021). For arguments against Contrastivism, see (Montminy & Russo, 2016; Steglich-Petersen, 2012).

[15] For the same reason as given by Schaffer (2016, p. 15), I leave off variations which take actual causation to be a *three*-part relation holding either between a cause, an effect, and a *causal* contrast class or between a cause, an effect, and an *effectual* contrast class. Such views seem to preclude the existence of casual chains.

(1) Susan stealing the bicycle caused her to be arrested.[16]

Proponents of a contrastivist view point out that whether (1) is true or false really depends on further details that have yet to be provided. Evaluating whether Susan's stealing the bicycle caused her to be arrested involves considering what would have happened had Susan *not* stolen the bicycle. But this is underspecified. What would she have done instead? Compare the following two ways of answering this question:

(1a) Susan stealing the bicycle *rather than the skis* caused her to be arrested rather than not be arrested.

(1b) Susan stealing the bicycle *rather than buying it* caused her to be arrested rather than not be arrested.

Intuitively, (1a) is false while (1b) is true. It is Susan's illegal action per se that causes her run-in with the law, not that she happened to steal a bicycle rather than some other piece of sporting equipment. Contrastivism preserves this intuition by holding that actual causation relates a cause and an effect *only* relative to a specification of contrasts for each. Different versions of Contrastivism result from different principles governing contrast selection. Northcott, for example, holds that contrasts must be *physically incompatible* with each other

[16] This example due to Dretske (1977), and is later picked up or adapted in (Hitchcock, 1996a, 1996b, 2011; Reiss, 2013a, 2013b)

– where the co-instantiation of any two contrasts would be inconsistent with the laws of nature – which roughly corresponds to the principle of exclusivity (Northcott, 2008, p. 118).

Arguably, Causal Relativism can be treated as one such version – a contrastivist view that names as the source of the contrasts the background space of possibilities. After all, Causal Relativism holds that actual causation relates a cause and an effect only relative to a specification of background possibilities, and such a specification then entails a set of contrasts for the cause and for the effect (and for all other features of the situation that are permitted to vary).

However, it merits emphasizing that while Causal Relativism can do the work done by Contrastivism, the converse does not seem to hold. I have argued that, in addition to the modal profile setting the contrast space, it plays a *direct role* in the evaluation of a causal claim. In **The Prince and His Biscuits**, specifying the causal contrast set as {Prince eats biscuits, Prince waters plant} and the effectual contrast set as {plant dies, plant survives} is yet *insufficient* to evaluate the causal claim in question. Whether the Prince eating biscuits caused the plant to die depends further on which modal profile is specified. I claim, therefore, that the modal profile is an *additional* parameter relative to which actual causation holds. And since the modal profile entails contrasts, it can arguably do the work of Contrastivism along the way. Of course, the viability of this last claim depends on how details of each view are worked out. The extent of this task unfortunately means that it lies beyond the scope of this dissertation, but I will make a few further remarks about this in Chapter 7.

4.6.d Realism and Pragmatics

This way of thinking about causation opens up what strikes me as a potentially extremely fruitful line of inquiry. In seeking to make explicit what has otherwise been a hidden parameter in causal claims, questions naturally arise about which modal profiles are of interest and why. Upon examination of everyday causal claims, for example, I suspect we will discover a preference for causal relations that are highly portable and robust, supporting accurate predictions and guiding successful behavior without requiring the careful tracking of background conditions. Relations of this sort will hold relative to those modal profiles that are constrained only by those contingent facts which commonly hold in everyday environments. Causal claims relative to modal profiles constrained by highly peculiar contingent facts will be unreliable unless such peculiar contingent facts are tracked, increasing cognitive load.

It is a major advantage of the proposed view that this preference can be explained in the obvious way – as due to the pragmatic benefit incurred. And yet, this invocation of pragmatic considerations in no way threatens the mind and language independence of causation. Once we fix on a modal profile, it is in no sense *up to us* what causes what. Instead, what is up to us is which of the many different possible real underlying causal structures we attend to.

This is precisely how this view can make room for the work done by the "serious" qualification on the principle of exhaustivity without undermining realism about actual causation, as I promised in Chapter 2. This version of exhaustivity requires that values of variables represent all *and only* those possibilities that we are willing to take seriously. This serves to rule out possibilities that are unlikely or irrelevant. Translated in terms of modal profiles, this qualification restricts the domain of eligible modal profiles – those that the SEM definition is taken to quantify over – to include only modal profiles that entail serious possibilities. Modal profiles that entail non-serious alternatives are excluded.

As mentioned earlier, however, such a restriction threatens to incorporate pragmatic considerations into the metaphysics of actual causation. Given Causal Relativism, though, we can permit this qualification on exhaustivity, treating it as an optional pragmatic operator that selects from the many different possible real causal structures those we might find particularly useful.

4.6.e Intrinsicality Principle of Variable Selection

I can now make good on the second promise I made in Chapter 2 – that this view can provide a satisfying error theory for the seemingly counterintuitive verdicts delivered by models that violate intrinsicality. To remind the reader, we considered the following example:

> **Plato's Grief** Socrates dies and Plato grieves the death of his teacher. But Plato has no fondness for Socrates's wife, Xanthippe, and would not be dismayed to see misfortune come her way.

Without intrinsicality, we are permitted to model **Plato's Grief** using a two-variable model interpreted so that one variable represents Xanthippe becoming a widow or not, and the other represents Plato grieving or not. But this model delivers the verdict that Xanthippe becoming a widow is an actual cause of Plato grieving. A prima facie counterintuitive result.

However, this can be made sense of when we take actual causation as relative to modal profile. Again, the intuition here strikes me as fuzzy. The causal verdict is not straightforwardly against intuition. There *is* a sense in which Xanthippe becoming a widow causes Plato to grieve, but there is also a sense in which this is not a cause of Plato's grief. Of course, Xanthippe becoming a widow causes Plato to grieve – given the background fact that *Socrates* is her husband. But it is not Xanthippe becoming a widow *simpliciter* that caused Plato to grieve. Had Xanthippe been married to someone else, as she very easily could have been, then her becoming a widow would not have caused Plato to grieve.

In terms of the view under discussion, Xanthippe becoming a widow is an actual cause of Plato grieving relative to the modal profile which is constrained by the fact that Socrates is Xanthippe's husband. But her becoming a widow is *not* an actual cause of Plato's grieving relative to any modal profile that relaxes this constraint. Xanthippe could have become a widow without Socrates dying – had she been married to anyone else, in fact – and had that happened then Plato would not have grieved.

Thus, we don’t need an intrinsicality principle – the relativity to modal profile can explain what feels counterintuitive about causal verdicts delivered by models representing non-intrinsic characterizations.

§4.7 Conclusion

I conclude that Causal Relativism about actual causation follows most naturally from the observation that models represent their target situations only relative to a modal profile. Of course, there are further ramifications of this observation to be explored, and further details about Causal Relativism to be filled in. I must leave much of this to future work. But the next chapter draws out one further advantage to this theory: if actual causation is relative to modal profile, then what counts as *proportional* will also be relative to modal profile. This, in turn, makes way for a defense of the principle of strong proportionality – the view that causation is essentially proportional.

CHAPTER 5

Strong Proportionality and Causal Claims

> … *x* can be causally sufficient for *y* even though it incorporates enormous amounts of causally extraneous detail, and it can be causally relevant to *y* even though it omits factors critical to *y*'s occurrence. What distinguishes causation from these other relations is that causes are expected to be *commensurate* with their effects: roughly, they should incorporate a good deal of causally important material but not too much that is causally unimportant.
>
> (Yablo, 1992, p. 273)

§5.0 Abstract There are several supposedly fatal objections to the view that causation is essentially proportional. This chapter shows how Causal Relativism can respond to all three in a unified way. I first articulate an amended proportionality principle, which I take to be an improvement on Yablo's original presentation of it. I then translate the amended principle into causal model terms. Finally, I argue that careful attention to how causal models represent, including the principles of variable selection and the relativity to modal profiles, dissolves the three objections from the literature.

§5.1 Introduction

Consider Sophie the pigeon who is trained to peck at and only at red things. Sophie pecks at a paint chip, which is a particular shade of red – scarlet.[1] Is the chip's being red or its being scarlet the cause of Sophie's pecking? Intuitively, *the* cause is the chip's being red. Had the chip not been red Sophie would not have pecked, whereas had it not been scarlet she still might have – had it been burgundy or crimson or some other shade of red.

This example illustrates the idea of a *proportional cause* – that cause that precisely made the difference. Very roughly, a cause will be proportional to its effect whenever its description includes enough but not too much causal information relevant to the description of that effect. In his (1992), Yablo suggests the following principle: that something is eligible to be called *the* cause of some effect just in case it is a proportional cause of that effect. Such a principle has been put to various philosophical uses: for example, as a proposed solution for the causal exclusion argument against non-reductive physicalism, and as a justification for and explanation of the privileging of higher-level causal explanations in the special sciences. However, the precise formulation of proportionality, and the significance it should be ascribed (if any), remains controversial.

The primary aim of this chapter is to defuse three objections to what can be called *strong* proportionality – the view that causation is essentially proportional, which I will clarify in

[1] This example is due to (Yablo, 1992). Notice that *Sophie* the pigeon pecks at red things, while *Alice* the pigeon, from Chapter 4, pecks at scarlet things. As noted, both examples come from the literature.

§5.2. One objection, put forward by Franklin-Hall, (2016), argues that the formulation of proportionality within a causal model framework is easily satisfied without successfully privileging intuitively proportional causes such as red in the Sophie example. It's therefore inadequate for capturing the kind of causal explanation we're looking for. The second objection, separately put forward by Bontly (2005), Weslake (2013), Franklin-Hall (2016), and McDonnell (2017, 2018), can be called *the problem of generic causes*.[2] It argues that proportionality legitimizes only the most general or abstract things as causes. Since many intuitive causal claims involve more specific causes than these, strong proportionality cannot be right. Finally, Shapiro and Sober (2012) demonstrate that proportionality counter-intuitively legitimizes disjunctive causes. This is particularly evident with things like non-monotonic functions.

My response to these objections employs the SEM framework. I first amend Yablo's original definition of proportionality (1992). I then propose a translation of the amended definition in terms of the SEM framework. The additional structure that a SEM framework introduces into the discussion plays a crucial role in my response – in particular, the articulation of a range of alternatives for each actual feature under consideration and the requirement that these ranges satisfy exclusivity, exhaustivity, and distinctness relative to a modal profile. My response then goes as follows.

First, a causal model formulation of strong proportionality works fine, contra Franklin-Hall, so long as the principle of exhaustivity is in place.

[2] Weslake (2013, p. 788) calls it the *problem of cheap sufficiency*.

Next, the problem of generic causes dissolves so long as exclusivity and exhaustivity are in place. By attending to the relativity of both exclusivity and exhaustivity to modal profile, it can be seen that the proffered examples of intuitive causes that purportedly fail to be proportional *are* proportional *relative to the implicit modal profile*. The objection that proportionality contravenes causal intuitions therefore doesn't go through.

Finally, I address the objection that proportionality counter-intuitively legitimizes disjunctive causes. I point out how much of what's compelling about this objection is dissolved when proportionality is amended in the way I suggest. However, something troubling remains. I argue that this is not as troubling as it seems.

§5.2 Strong versus Weak Proportionality

The task of articulating a notion of proportional causation requires first getting clear on the work it is meant to do. I will therefore begin by delimiting several positions one could take on the significance of proportionality – of whether proportional causes are privileged in some way. First, of course, one could take a skeptical position – denying either the value or coherence of proportional causation altogether.[3] This view is often the result of first consideration and then rejection of two other possible positions: *weak* proportionality and *strong* proportionality.

[3] For proportionality skeptics, see (Franklin-Hall, 2016; Hoffmann-Kolss, 2014; Menzies, 2008; Shapiro & Sober, 2012).

The position of *weak proportionality* treats proportionality as a merely optional pragmatic constraint on causal explanations.[4] Weak proportionality holds that proportional causes are at least sometimes better to cite than non-proportional ones for pragmatic reasons. For example, it may be the case that a cause satisfying the proportionality conditions is more explanatorily useful in some communicative context than any failing to so satisfy. As a question of pure semantics, though, proportional and non-proportional causes are equally eligible to be called 'the cause' of some effect.

Proponents of strong proportionality disagree. *Strong proportionality* holds that one thing can be called *the* cause of something else only if it is proportional.[5] What is at issue in the debate over strong proportionality is a claim about natural language – that claims of the form '*c* is the cause of *e*' are true only if *c* is a *proportional* cause of *e*. This is the view that I will defend.

One clarification of this view is worth emphasizing. This is that strong proportionality does not – or, at least, need not – hold that proportional causation is the *only* causal relation. It would be hard to deny the existence of various other relations of causal influence aside from proportional causation. For example, factors can be causally *sufficient* and causally *relevant* for a given effect, despite their not being proportional. A factor is causally sufficient by being enough for the effect to manifest without being necessary, and a factor is causally relevant

[4] For proponents of weak proportionality, see (Blanchard, 2018; Bontly, 2005; Maslen, 2017; McDonnell, 2017, 2018; Weslake, 2013, 2017; Woodward, 2015; Yablo, 2003).

[5] For proponents of strong proportionality, see (List & Menzies, 2009; Menzies & List, 2010; Papineau, 2013; Yablo, 1992).

by being necessary for the effect to manifest without being sufficient. It is then open to a strong proportionalist to allow for our causal talk picking out, at different times, any of these other relations of causal influence, in addition to proportional causation. Plausibly, the terms 'causes' and 'caused' are ambiguous in natural language between the various relations of causal influence. Strong proportionality is a view *only* about the meaning of claims of the form "*c* is *the* cause of *e*." It does not apply to all causal claims whatsoever. So, claims of the form "*c* causes *e*" can be equally taken to mean that *c* and *e* stand in the proportional causation relation, causal sufficiency relation, causal relevance relation, etc. The relevant intuitions against which to check the plausibility of strong proportionality are therefore those about claims of the form 'This is *the* cause of that,' rather than about claims of the form 'This *causes* that.' In line with this, I should also point out that it need not be a commitment of the strong proportionalist that causal models must represent proportional relata. Causal models can still be used to model merely causally sufficient and/or causally relevant relata.

Now that we have its purpose on the table, I can turn to the task of articulating what it is for a cause to be proportional.

§5.3 Articulating Proportionality

The definition of proportionality that I endorse borrows heavily from Yablo (1992, pp. 273–277). Yablo defines proportionality using four counterfactual conditions – Contingency, Adequacy, Required, and Enough – and a cause is proportional to its effect just in case it satisfies all four of these conditions. My proposal will amend this by dropping Required and revising Enough.

5.3.a Contingency + Adequacy

The first two conditions of Contingency and Adequacy serve first to ensure that the proportional cause is in fact a cause.[6] Contingency captures the idea that the absence of a cause results in the absence of the effect. It requires that were c not the case, then e would not have been, or, $\neg c \mathrel{\Box\!\!\rightarrow} \neg e$. Adequacy captures the idea that the effect would have occurred had the cause occurred. It requires that were c not the case then, had it been, then e would have been the case, or, $\neg c \mathrel{\Box\!\!\rightarrow} (c \mathrel{\Box\!\!\rightarrow} e)$.[7]

Contingency + Adequacy, taken together, also serve to capture some of the proportionality intuition. Contingency blocks the instantiation of overly specific properties from counting as commensurate. For example, say Socrates drinks hemlock and dies. In fact, he *guzzles* it. But suppose that his *drinking* hemlock was sufficient for his dying. Then, was Socrates's *drinking* or *guzzling* hemlock proportional to his dying? Contingency + Adequacy say that it was his *drinking* hemlock, since his *guzzling* fails to satisfy Contingency. Had he not guzzled the hemlock, but still drank it, then he still would have died.

[6] Contingency and Adequacy on their own are not enough to handle redundant causation, of course. But the discussion here merely serves to inform what proportionality should look like in the causal model framework, in which redundant causation is handled by AC2 of $\boldsymbol{AC - relative}$ (or something like it).

[7] Adequacy updates the simpler counterfactual condition: had c been the case then e would have been the case; $c \mathrel{\Box\!\!\rightarrow} e$. But this simpler condition is trivially satisfied whenever c and e are actually the case. The updated version results in a non-trivial condition in cases where both c and e are the case.

Adequacy blocks the instantiation of overly general properties from counting as commensurate. For example, say that Xanthippe shuts the door and it falls off its hinges. In fact, she *slams* it. But suppose that her slamming the door was necessary for it to fall off its hinges. Then, was Xanthippe's *shutting* or *slamming* the door proportional to it falling? Contingency + Adequacy say that it was her *slamming* the door, since her *shutting* it does not satisfy Adequacy. Had she not shut the door, then had she shut it but not slammed it, it would not have fallen off its hinges.

However, Contingency + Adequacy do not fully capture the proportionality intuition. To see where they fall short, consider the following.

> **Socrates's Sloppy Habits** Socrates is a sloppy drinker. Due to an esophageal abnormality, he cannot drink anything without guzzling it. He guzzles hemlock and dies.[8]

This draws our attention first to the fact that, in some cases, multiple causes at different levels of description may satisfy Contingency + Adequacy. In **Socrates's Sloppy Habits**, they are satisfied by both Socrates's *guzzling* the hemlock and by his *drinking* it. His dying becomes contingent on his guzzling hemlock in **Socrates's Sloppy Habits** due to his inability to drink without guzzling. Had Socrates not guzzled hemlock, then he would not have (indeed, *could* not have) drunk hemlock, and so he would not have died. His guzzling is also adequate for his dying since, had he not guzzled then, had he guzzled, he would have died. But Socrates's *drinking* hemlock satisfies Contingency + Adequacy, as well. His dying is

[8] This example is adapted from (Yablo, 1992, p. 276).

contingent on his drinking since had he not drunk hemlock, then he would not have died. His drinking is also adequate for his dying since had he not drunk then, had he drunk, he would have died. This is no good. An account of proportionality should identify what is *uniquely* commensurate with the effect. Additional conditions are called for.

5.3.b Two New Conditions: Required and Enough

In response, Yablo introduces two new conditions: Required and Enough. He motivates them with the intuition that, even in **Socrates's Sloppy Habits**, Socrates's *drinking* hemlock is a better candidate for being the proportional cause than his *guzzling* it. Yablo points out that even though it is not possible that Socrates drink anything without guzzling it, we can still ask what would have happened had he done so. And had he drunk the hemlock without guzzling it, then he still would have died. The condition of Required is what permits consideration of what would have happened had Socrates's *drinking* hemlock taken place without his *guzzling* hemlock, despite the fact that this is an in fact remote possibility.

Required For all $c^{-} < c$, if c^{-} had occurred without c, then e would not have occurred.[9]

With Required in place, Socrates's drinking qualifies as proportional over his guzzling, since if his drinking had occurred without his guzzling, then he still would have died.

[9] A natural way to read $<c^{-}, c, c^{+}>$ assumes that the properties that partially comprise each property instance, c^{-}, c, c^{+}, stand in a determinable-determinate relation where c^{-} is the least and c^{+} is the most specific. For example: *being-colored*, *being-red*, *being-scarlet*. But it is not essential that the instantiated properties stand in a determinable-determinate relation. The essential requirement is merely that the property that partially comprises c^{+} is a more specific way that the property that partially comprises c could have been instantiated, and the same of c relative to c^{-}.

Required covers any situation where, due to constrained background circumstances, an overly *specific* property being instantiated satisfies Contingency + Adequacy. The second additional condition, Enough works in the other direction, covering cases where constrained background circumstances make it so that an overly *general* property being instantiated satisfies Contingency + Adequacy.

Enough For all $c^{+} > c$, c^{+} was not required for *e*.

As an example of the kind of case that Enough is designed to handle, consider:

> **Xanthippe's Oiled Door** Xanthippe puts way too much oil on the hinges of her door. As a result, the door cannot be shut without it being slammed. Later, she slams the door, and it flies off its hinges.

Normally, one could shut a door without slamming it and so, as we saw above, Xanthippe's shutting the door would not be adequate for it falling off its hinges. But in **Xanthippe's Oiled Door**, her shutting the door becomes adequate for it flying off its hinges due to it having been excessively oiled. Had Xanthippe not shut the door, then had she shut it, she would have (indeed, *could only* have) slammed it, and so it would have flown off its hinges. The condition of Enough permits consideration of what would have happened had Xanthippe *shut* the door without *slamming* it, despite the fact that this is an in fact remote possibility. With Enough in place, Xanthippe's *slamming* the door qualifies as proportional over her merely *shutting* it, since her *slamming* the door is required for it to fall off its hinges.

So, the addition of these two conditions succeeds in privileging Socrates's *drinking* hemlock over his *guzzling* it as a more commensurate cause of his dying, and Xanthippe's *slamming* the door over her *shutting* it as a more commensurate cause of the door flying off its hinges. However, as many have pointed out, an account of proportionality defined by Contingency + Adequacy + Required + Enough takes things too far.[10] For example, while they privilege Socrates's *drinking* over his *guzzling*, they do not privilege his *drinking* full stop. His *drinking* does not satisfy Required – had he merely consumed the hemlock without drinking it, he still would have died. Thus, they privilege Socrates's *consuming* hemlock over his *drinking* it. But it goes further. His *consuming* fails to satisfy Required, as well – had he (*consumed or injected*) hemlock without consuming it, he still would have died. Thus, the truly commensurate cause that satisfies these conditions turns out to be only the instantiation of some very general or abstract property.

5.3.c Causal Relativism and Enough*

A causal relativist about actual causation can explain where this account goes wrong. By hypothesis, the possibilities originally presupposed by **Socrates's Sloppy Habits** are such that Socrates cannot drink without guzzling. When we ask anyway what would have

[10] This is what I'm calling *the problem of generic causes*, which I will discuss in greater detail in §5.5. See Bontly (2005), Weslake (2013), Franklin-Hall (2016), and McDonnell (2017, 2018).

happened had Socrates drunk hemlock without guzzling it, we shift the space of possibilities under consideration. Similarly, by hypothesis, the possibilities originally presupposed by **Xanthippe's Oiled Door** are such that the door cannot be shut without being slammed. When we ask anyway what would have happened had Xanthippe shut the door without slamming it, we shift the space of possibilities under consideration. But what justifies these shifts? Why should we privilege what is commensurate relative to the more permissive space of possibilities over what is commensurate relative to the more constrained space of possibilities?

One answer that comes to mind is that perhaps Required and Enough discount what is commensurate relative to any one particular space of possibilities – such as that implicit in **Socrates's Sloppy Habits** – in order to identify what is commensurate in a sense that transcends this relativity. Such an account aims to capture an idea of proportional causation simpliciter. But a causal relativist has grounds to question this aim. According to causal relativism, actual causation already holds relative to a modal profile. The natural view would therefore take *proportional* causation to also hold relative to modal profile.[11] On this view, a particular cause is proportional to a particular effect *only* relative to a modal profile. Strong proportionality would then hold that a claim of the form '*c* is *the* cause of *e*' picks out the proportional cause relative to the implicit modal profile. As I will argue, this view can respond well to several objections against strong proportionality.

[11] Logically speaking, a causal relativist about actual causation could introduce a notion of proportionality that eschews relativity in some way – perhaps by taking a privileged modal profile or by somehow quantifying over all modal profiles or over some privileged subset. I'm not sure how coherent a notion could be achieved, nor how useful it would be in the end. Regardless, I suspect it would be a pragmatic notion, for similar reasons as discussed in §4.5.b.

Taking proportional causation to hold relative to modal profile renders Required and Enough inappropriate. But without Required and Enough, we're back to the original issue – where more than one cause may qualify as proportional. Relative to the implicit modal profile, both Socrates's *drinking* and his *guzzling* hemlock satisfy Contingency + Adequacy in **Socrates's Sloppy Habits**, and both Xanthippe's *shutting* and her *slamming* the door satisfy Contingency + Adequacy in **Xanthippe's Oiled Door**. This is because, within this modal profile, Socrates will instantiate the property of drinking just in case he instantiates the property of guzzling. The instantiations of these two properties by Socrates are co-extensive within this modal profile. But <Socrates, drinking> and <Socrates, guzzling> involve different properties, and so are distinct. We therefore get two different, albeit modally co-extensive, property instantiations as proportional causes. The same can be said of Xanthippe shutting and her slamming of the door. Something else is still needed for proportionality to deliver a uniquely commensurate cause.

In response, I propose that the Contingency + Adequacy account be supplemented with a condition satisfied only by the instantiation of the most specific property that also satisfies contingency. I call this Enough*.

Enough* For all $c^{+} > c$, e is not contingent on c^{+}

Contingency + Adequacy + Enough* endorses Socrates's *guzzling*, and *only* his guzzling, as the proportional cause in **Socrates Sloppy Habits**, relative to the modal profile implicit in the case – namely, given Socrates's esophageal abnormality. Socrates's *drinking* fails to satisfy Enough*, since the effect of Socrates dying was contingent on the instantiation of a more specific property – namely, his *guzzling*. Had his *guzzling* not occurred, then his death would not have occurred. But the buck stops here. There is no property more specific than his guzzling on which Socrates's dying was also contingent. Socrates could have failed to guzzle in the precise way he in fact did, for example, but still would have died had he guzzled in some other way.

This may strike the reader as a strange result. Indeed, it was Socrates's guzzling being proportional that struck Yablo as counterintuitive and motivated his introduction of Required and Enough. It may be worth emphasizing, then, that this result is not the same as before. The causal relativist can agree that it is counterintuitive to say that Socrates's *guzzling* is the proportional cause *simpliciter*. But that isn't what's being said here. On a causal relativist view, there is no such thing as the proportional cause simpliciter. Socrates's guzzling is proportional to his dying *given the fact that Socrates is incapable of drinking without guzzling*. Is this result so strange? Socrates's guzzling *is* the most precise difference maker in this context. His guzzling is the most specific target upon which an intervention would toggle the effect in this situation. Change his guzzling, change his dying. Were we to consider **Socrates Sloppy Habits** relative to a modal profile where Socrates *could* drink without guzzling, then his guzzling would no longer satisfy Contingency. In that case, Socrates's *drinking* would be the proportional cause *given the fact that Socrates, in some sense,* could *drink without guzzling*.

Similarly, Contingency + Adequacy + Enough* endorses Xanthippe's *slamming* the door, and only her slamming, as the proportional cause in **Xanthippe's Oiled Door**, relative to the modal profile implicit in the case – namely, given the excessive oil. Xanthippe's merely *shutting* the door fails to satisfy Enough*, since the effect of the door flying off its hinges is contingent on the instantiation of a more specific property – namely, her *slamming* the door.

I hereby endorse an account of proportionality that combines Adequacy, Contingency, and Enough*. A cause is *proportional* to an effect relative to a modal profile only if it satisfies all three conditions relative to that effect and that modal profile.

> **Proportionality** A property instantiation, c, is the *proportional* actual cause of a property instantiation, e, relative to a modal profile, θ_i, just in case c satisfies Contingency, Adequacy, and Enough* relative to e and θ_i.

§5.4 Translating Proportionality into Causal Model Terms

Now to translate this idea of proportionality into causal model terms. Models introduce additional structure by requiring that a range of mutually exclusive and jointly exhaustive alternatives be articulated for each actual, distinct feature under consideration. Further, as I've argued, ranges satisfy exclusivity, exhaustivity, and distinctness only relative to a modal profile. My response to the three objections to strong proportionality takes advantage of this additional structure, relying on each of exclusivity, exhaustivity, distinctness, and relativity to modal profile. Since no precise causal model definition of proportionality exists in the literature, the following will be my own presentation.

5.4.a Defining Proportional Variable

I propose that proportionality between a cause and an effect be defined partly in terms of proportionality between the *variables* that represent them. Thus, a cause and effect will be proportional only if they can be represented by proportional variables in an apt model. As a first pass, two variables will be *proportional* just in case changes in one of them (the cause variable) line up in the right way with changes in the other (the effect variable). This captures the intuition that the paint chip's being red is proportional to Sophie's pecking because changes in the chip's shade from red to otherwise will correspond to changes in whether Sophie pecks. But more still needs to be said about what it is to 'line up in the right way'.

To motivate the definition I eventually adopt, I'll first translate Yablo's example into causal model terms. Take the variable, P, to be a variable representing whether Sophie the pigeon pecks or not. It can take the values: {*peck*, *not-peck*}. Now consider two alternative variables for representing the property-instantiations of the paint chip: the variable, R, which can take the values {*red*, *not-red*}, and the variable, T, which can take the values {*taupe, scarlet, cyan, mauve, crimson,* etc.}, where 'etc.' stands for all other physically possible colors at the same grain as those already made explicit. These two variables are equally good in a number of ways. First, variables R and T each respect the exclusivity and exhaustivity principles. Second, each of R and T counts as causally related to P according to the core criterion of causation given above. The variable P will change values whether you intervene to change the value of R from *red* to *not-red*, or intervene to change the value of T from *crimson* to *taupe*. But the relationships that R and T respectively bear to P are different. *All* of the changes in R line up with changes in P – every intervention on R corresponds to P taking a different value. There is a one-to-one correspondence between the values of R and those of P. But only *some* of the changes in T line up with those in P – only certain interventions on T correspond to P taking a different value. If the value of T is *taupe*, say, then the intervention that assigns T the value *cyan*, for example, does not so correspond. So, there is not a one-to-one correspondence between the values of T and P.

This is the feature we're looking for to define proportionality. Variable R is *proportional* to variable P, while T is not, because the values of R counterfactually line up one-to-one with those of P, while those of T do not. Two variables will line up in this way whenever it's the case that every intervention on one variable, setting it to a new value, leads to a change in value in the other variable. More precisely:

> **Proportional Variable** X is a *proportional variable* relative to Y, given a model $\mathcal{M}_i$, just in case every intervention on X, and on X alone, leads to Y taking a different value.[12]

A given cause, c, will then be *proportional* to an effect, e, only if there is an apt model in which c can be represented by a proportional variable relative to e. However, this does not yet

[12] This is perhaps what Woodward means to pick out with his Principle P (2010, p. 298), but perhaps not. His principle is too vague to tell.

incorporate the insight from §5.3. Both Socrates's drinking and his guzzling would still qualify as a proportional cause of his dying. We also need to translate Enough* into causal model terms, so as to ensure a unique proportional cause even in constrained circumstances. I propose the following translation.

> **Enough-CM** For all $c^{+} > c$, there is no apt model that represents c^{+} by a proportional variable relative to *e*.

Putting this together, we get the following causal model definition of proportionality.

> **Proportionality-CM (*P-CM*)** A property instantiation, *c*, is a *proportional* actual cause of a property instantiation, *e*, just in case there is an apt model $\mathcal{M}_i$, according to which *c* is represented by *X* = *x*, and *e* by *Y* = *y*, <*X* = *x*, *Y* = *y*> satisfies ***AC – relative*** in $\mathcal{M}_i$, *X* is a proportional variable relative to *Y*, and *c* is Enough-CM for *e*.[13]

5.4.b Proportionality Does the Trick

With a causal model definition of proportionality in place, I can now discharge Franklin-Hall's objection (2016). Franklin-Hall contends that no formulation of proportionality in terms of causal models can successfully prioritize intuitively proportional causal relata, such as the chip's *being red* in the Sophie the pigeon example.

[13] The definition specifies that *there is* an apt model, rather than *for every* model, to allow for the possibility that there may be an apt model in which the proportional cause doesn't even appear.

Refer back to Sophie and her paint chip. Franklin-Hall introduces a comparison between the variable, *R*, that can take the values: {*red*, *not-red*}, (as above), and the variable, *C*, that can instead take the values: {*cyan, scarlet*} (as above). *R*, as before, is proportional to, and therefore a genuine cause of, *P*. But, she argues, *C*, too, is proportional to *P*, since *C* seems to be a proportional variable relative to *P*. An intervention on *C* that changes its value from *cyan* to *scarlet* changes *P* from *not-peck* to *peck*, and an intervention that changes *C*'s value from *scarlet* to *cyan* changes *P*'s value from *peck* to *not-peck*. Thus, the changes in *C* line up with the changes in *Y* just as well as the changes in *R* do. The problem, then, is that ***P-CM***, as formulated, is insufficient to its intended task. It fails to privilege a variable like *R* over one like *C*, and so fails to prioritize a causal model that uses *R* over one that uses *C*.

In response to this problem, a natural move would be to find a way to disqualify variables like *C* from the arena. Intuitively, *C* is not the right kind of variable. But, why not? Simple. Our aversion to variables like *C* is due to their failure to satisfy exhaustivity relative to the implicit modal profile of the situation.[14] The paint chip could have been any physically possible color. Unless the possible color of the paint chip is restricted in some special way – by a local factory, perhaps – then the underlying object could fail to take one of *C*'s two values. There are other possible colors that the paint chip could have had – such as beige or olive green. In failing to represent these possibilities, *C* fails to satisfy exhaustivity. Thus, *C* is an inapt

[14] Note that this is *not* the notion of exhaustivity discussed in (Franklin-Hall, 2016). Franklin-Hall proposes and dismisses an exhaustivity condition that "requires that the cause variable's values collectively exhaust the…range of circumstances by which the explanandum event – as well as its contrast – might be brought about." (2016, p. 566)

variable, and any model constructed using C is an inapt model. This means that ***P-CM*** goes unsatisfied, despite C being a proportional variable.

The variable, R, on the other hand, is exhaustive, since the object must take one of R's two values. So, C is discounted as a candidate variable relative to the implicit modal profile, and so *scarlet* is disqualified as the proportional cause. R remains as a proportional variable, and so *red* is the proportional cause of Sophie pecking.

In general, two cause variables compete over which is proportional to some effect variable only when they are exhaustive relative to the same modal profile. C and R would only be competitors for proportionality relative to a modal profile according to which the only possible colors of the paint chip were scarlet and cyan. If this were the modal profile in the background – if we were in the factory yard, as in **Alice in the Factory** from §4.2.a, for example – then both C and R would be exhaustive variables, and both would be eligible to figure in an apt model. Only in such a case would C and R compete for proportionality. And due to Enough-CM, the outcome of the competition would be that *scarlet* is proportional to Sophie pecking – proportional given the fact that we're in the factory yard, that is.

§5.5 The Problem of Generic Causes

I next turn to what I take to be the primary objection pitched against strong proportionality – what I call *the problem of generic causes*. This objection relies on the argument that only very general or abstract causes can satisfy proportionality. Of course, many things that we would naturally call *the* cause of some effect are not general in this way. As a claim about natural language, then, strong proportionality can't be right. (Bontly, 2005; Franklin-Hall, 2016; McDonnell, 2017, 2018; Weslake, 2013).

5.5.a The Objection

To illustrate, I'll work with one of Bontly's examples, but Weslake's and McDonnell's are of the same kind. Take a simple case where Socrates drinks hemlock and then dies, and the corresponding causal claim, 'Socrates's drinking hemlock is the cause of him dying'. (We are no longer imagining Socrates with an esophageal abnormality.) This claim sounds right. But the objection is that drinking hemlock is not actually proportional to Socrates dying. If Socrates had not drank hemlock but still consumed it – by eating a dozen leaves, perhaps – then he still would have died. It seems that his drinking hemlock therefore fails to satisfy Contingency. In causal model terms, this seems to show that the changes in the variable that would represent Socrates's drinking hemlock won't line up in the right way with the changes in the variable that would represent Socrates's dying. An intervention on the first variable could change its value from *having-drank-hemlock* to *having-eaten-hemlock* and the second variable would *not* change its value, retaining the value *dies*. This causal claim is therefore not proportional. The proportional cause should be, instead, Socrates's consuming hemlock.

The objection can be run even further. The above would need to assume that no other lethal forces are at play. But Socrates also would have died had he performed seppuku, or had he refused food and drink for several days. So, the real proportional cause should be, instead, something like *having-had-a-lethal-experience*. For only something as general or abstract as this could genuinely satisfy Contingency, and therefore proportionality.

In simple terms, the response from a causal relativist is that each introduction of an even-more-remote possibility changes the modal profile relative to which the proportional cause needs to hold. The objection equivocates on background modal profile. But it also fails to attend to what counts as a genuine range of alternatives. Socrates's drinking and eating hemlock are not mutually exclusive relative to all modal profiles. This response is easiest to see once the examples are translated into the causal model framework. Once you set up a model in a way that respects exhaustivity, exclusivity, and distinctness relative to a given modal profile, the problem dissolves.

5.5.b How Exhaustivity Preserves Causal Intuitions

Take the hemlock example just outlined. Importantly, this example and corresponding claim are under-defined.[15] Translated into causal model terms, all that this description provides is that there is some variable that takes at least one value that represents Socrates's drinking hemlock, and an intervention on this variable changes the value of some other variable to one that represents Socrates's dying. But, a number of different variables could represent the purported cause, and a number of different models could represent its relationship to the effect of Socrates' dying. Which of these is representationally accurate depends on what the relevant alternatives to drinking hemlock are – i.e. what the background modal profile is.

[15] This is not a new observation. See (Franklin-Hall, 2016; McDonnell, 2017; Weslake, 2017)

How these details get filled in will also determine whether or not the variable that represents Socrates's drinking hemlock is a proportional variable, and so whether his drinking hemlock is a proportional cause of his dying.

I hold that the implicit modal profile in the vignette is that hemlock was the only possible poison, and drinking it the only possible means of consumption, for reasons I'll provide in §5.5.c. Given this, the exhaustive variable representing Socrates's drinking hemlock has the values {*having-drank-hemlock*, *not-having-drank-hemlock*} – call this *A*. The exhaustive variable representing Socrates's death has the values {*having-died*, *not-having-died*} – call this *Q*. But, *A* is indeed proportional to *Q*. When an intervention sets the value of *A* to *having-drank-hemlock*, *Q* takes the value *having-died*. When an intervention sets the value of *A* instead to *not-having-drank-hemlock*, *Q* changes value to *not-having-died* (since there's no other way for Socrates to consume poison). Socrates's drinking hemlock is, furthermore, the most specific property instance that can be represented by a proportional variable in an apt model relative to this modal profile, thus satisfying Enough-CM. So, the intuitive cause of Socrates's drinking hemlock is proportional after all.

5.5.c Implicit Modal Profiles

This response first requires that causal claims be implicitly relative to a modal profile, as I argued at the end of Chapter 4. However, any kind of relativity of a causal claim is explicitly denied by both McDonnell and Weslake (McDonnell, 2017; Weslake, 2017). They each claim that if causal claims are indeed relative in this way, then we wouldn't be able to agree on the truth value of the claim without first settling on the profile. They argue that the very fact that we have strong and convergent intuitions about these examples, despite their being underdetermined, demonstrates that the intuitions are not sensitive to filling in modal details.

In response, I concede that our having strong and convergent intuitions about vignettes indicates that we must implicitly agree on a modal profile. However, I would argue that this is precisely what we do – we implicitly agree on a modal profile. Moreover, we achieve this agreement in the very same, non-mysterious way that we naturally achieve agreement about the plethora of missing context in everyday conversations.[16] My preferred explanation of this remarkable yet plebeian phenomenon comes from Grice. According to Grice, communication generally is governed by a set of unspoken but presupposed conversational maxims (1989). The maxims most relevant here are those of *quantity* and *relation*. Taken together, these maxims enjoin an interlocutor to,

> Make your contribution as informative as is required (for the current purposes of exchange)….[and no] more informative than is required,….[and b]e relevant. (1989, pp. 26–27)

The principle of charity has us assume that the presenter of a vignette respects these maxims. Thus, the natural way to fill in the modal profile of these examples is to take each fact as informative and relevant, and to assume that all informative facts have been provided.

[16] A similar point is made, to different effect, by Bontly (2005). It may also need saying that this move bears resemblance to that made by any view that takes causal claims to be sensitive to contrasts, where the contrasts are set by conversational context (Schaffer, 2012; Shapiro & Sober, 2012).

My defense further requires that the implicit modal profile is as I've specified. My reasoning follows. The only information provided by the example is that (i) Socrates drinks hemlock; and (ii) Socrates dies. Assuming that this is all the information we explicitly need to be given to understand what's going on, and thus that nothing significant has been left out, any unspecified details should be filled in consistent with everyday life. The vignette tells us that Socrates drank hemlock. This is not a normal thing to drink, nor to consume in any way. Few people have experience with hemlock consumption in general. The little exposure one might have – especially as a philosopher – is to this exact story of Socrates drinking hemlock as a form of execution and subsequently dying. Alternative ways of consuming this poison don't arise. Further, there is nothing in the vignette to suggest that there are alternative means of consuming the hemlock. Thus, treating his *eating* hemlock, for example, as a relevant alternative would be to arbitrarily introduce something that wasn't otherwise specified, and whose presence can't be justified by everyday experience. A similar story can be told about the ingestion of poison of any kind. As a result, the only real alternative to Socrates's drinking hemlock is his not consuming poison.

The problem of generic causes seems to get off the ground because it stipulates what seems like a range of relevant alternative possibilities to Socrates's drinking hemlock, and then argues that given these other possible alternatives, the causal claim is not proportional. However, I have argued that the intuitive cause in this case is implicitly relative to a modal profile that doesn't include these other alternatives. To introduce these other alternatives is to introduce a different background than what is implicitly in play, and thereby to change the subject. Relative to the modal profile that I take to be implicit, the intuitive cause is proportional.

5.5.d How Exclusivity and Distinctness Preserve Causal Intuitions

But let's say that the alternatives introduced by the objectors *are* live possibilities. That is, let us assume that the implicit modal profile is whatever would be required for the alternatives introduced in the objection to be relevant. Even so, the intuitive cause still comes out proportional. This is due to the principles of exclusivity and distinctness.[17] The problem presupposes that there is some relevant alternative to Socrates's drinking hemlock that preserves his consuming it. Take as an arbitrary alternative his eating hemlock. But Socrates could both drink and eat the hemlock – he could wash down a hemlock salad with a full glass of hemlock milk, for example. Since these possibilities can occur together, exclusivity and distinctness dictate that they should be represented by different variables. For example, let's represent them using two variables – *B:* {*having-eaten-hemlock*, *having-not-eaten-hemlock*} and *D:* {*having-drank-hemlock*, *not-having-drank-hemlock*}. There is still no problem here for strong proportionality. D is a proportional variable to Q, and there is no more specific property instance that can be represented by a proportional variable. So, Socrates's drinking the hemlock is again the proportional cause.

There is, however, a way to manufacture a proportionality problem. Imagine that Socrates's jailor only has enough money to purchase either hemlock leaves or hemlock milk, but not

[17] This is a similar move as that made in (Woodward, 2018).

both. If we take the jailor's budget as a fixed part of the background conditions, then Socrates's options become constrained so that his drinking hemlock excludes his eating it, and vice versa. Thus, the property instantiations of Socrates *having-drank-hemlock* and *having-eaten-hemlock* should be values of the same variable. Call this variable *H:* {*having-drank-hemlock, having-eaten-hemlock, having-neither-drank-nor-eaten-hemlock*}*.* *H* is not a proportional variable to *Q*, since an intervention on *H* that changes its value from *having-drank-hemlock* to *having-eaten-hemlock* will not correspond to a change in *Q*. Thus, neither Socrates's drinking nor his eating will be proportional. The proportional variable would instead be one that has the values {*having-consumed-hemlock*, *not-having-consumed-hemlock*}. So, the proportional cause of his dying is instead his consuming hemlock.

As always, the proportional cause in this instance holds relative to a modal profile. Socrates's consuming hemlock is not the proportional cause *simpliciter* (since there isn't one), but the proportional cause of his dying *given the jailor's budget*. This is not so strange. After all, it isn't the drinking in particular nor the eating in particular that makes a difference to whether Socrates dies, since a salient reason for him not doing one is that he in fact did the other. What makes a difference in this situation is whether he *consumes* hemlock, regardless of whether he drinks or eats it.

§5.6 The Problem of Disjunctive Causes

Shapiro and Sober (2012) raise a similar objection – that strong proportionality will delegitimize many intuitive causal claims. But their reasoning behind this merits its own response. They draw our attention to the case of disjunctive causes, presenting an example of a non-monotonic function, $f(X) = Y$, in which both $X = 3$ and $X = 22$ will produce the same effect of $Y = 6$. In such a case, neither $X = 3$ nor $X = 22$ is a proportional cause of $Y = 6$. Assume a situation in which $X = 3$, and so $Y = 6$. But had it instead been the case that $X = 22$, it would still be the case that $Y = 6$. Thus, $X = 3$ is not a proportional cause. The same can be said of $X = 22$. This is one instance of the general phenomenon of some effect being caused by two different things. The truly proportional cause in these cases seems like it must be a disjunction. In this case, the disjunction of being 3 or being 22.

5.6.a An Overstated Case: "the" Cause and Inclusive Disjunctions

Shapiro and Sober conclude that strong proportionality "will mean rejecting almost all the causal statements we think are true." (Shapiro & Sober, 2012, p. 90).[18] But this is too quick for two reasons. First, they fail to distinguish between claims of the form "*c causes e*" and those of the form "*c* is *the* cause of *e*," and systematically employ the former type rather than the latter in their examples. As a result, they overstate their case. Strong proportionality applies only to claims of the latter form. We can all still agree that '$X = 3$ causes $Y = 6$' without thereby committing ourselves to the claim '$X = 3$ is *the* cause of $Y = 22$'. After all, X taking the value *3* or X taking the value 22 can still be *causally relevant* to Y taking the value *6* in some situation, despite it not being *proportional* to that effect.

[18] The fatality of this problem is agreed to in (Weslake, 2017; Woodward, 2018)

Second, arguably the kind of disjunctive causes that would be responsible for contravening "almost all" of our causal intuitions are those that disjoin independent properties. These take a form similar to the disjunctive cause in the following claim: "The cause of Sophie's pecking is that '[she was] presented with any red target, [*or* she was] provided food, [*or* she was] tickled, and so on (Franklin-Hall, 2016, pp. 566–577).'" Were this cause to be represented by a single variable, with a different value mapping to each of *being-presented-with-a-red-target*, *being-provided-food*, *being-tickled*, and etc. Since these property instances are *not* mutually exclusive, however, such an interpreted variable would violate exclusivity and therefore be impermissible. The target's being red and Sophie's being tickled should be values of different variables. This is simply the same argument from before, in §5.5.d.

One may object that there's nothing to stop us from modeling the *disjunction itself* as the value of a variable. We would then have a variable that represents by one of its values the disjunctive property instantiation of *being-presented-with-a-red-target*-or-*being-provided-food*-or-*being-tickled*, and perhaps the negation of this by its other value. The result is that *this* would be a proportional variable to *P* – the variable that represents Sophie pecking. But this objection goes no further. This is *a* proportional variable, but it is not the finest-grain proportional variable. Enough-CM is not satisfied. There is a finer grained property instantiation – *being-presented-with-a-red-target* – that can also be represented by a proportional variable in an apt model-interpretation pair - namely, the variable *R* which represents Sophie being presented with a red target when it takes the value *1* and Sophie not being presented with a red target when it takes the value *0*. It is still the case, then, that the chip's being red is the proportional cause.

5.6.b Exclusive Disjunctions

The problem has been whittled down, but it has not been removed. There *is* a kind of disjunction that seems to pose an actual threat to strong proportionality – the kind where the disjuncts are mutually exclusive properties. For example, if Sophie pecked at all and only blue or red things. Proportionality would dictate that the cause of Sophie's pecking would therefore be the disjunctive property of the chip's *being-red-or-blue*.

Arguably, the best response on behalf of strong proportionality is indeed to bite the bullet and accept disjunctive causes of this kind. But it is worth flagging that this is not so unpalatable in many cases. Consider again the final case from §5.5.d where, due to the constrained circumstances, the possible instantiations of the properties *having-eaten-hemlock* and *having-drank-hemlock* are mutually exclusive. As a result, the proportional cause is the disjunctive property instantiation Socrates's *having-eaten-or-drank-hemlock*, or, in other words. the property instantiation of Socrates *having-consumed-hemlock*. In cases like these, there is an appropriate single term for the relevant disjunction.

Sometimes, however, there is no neat, single term. This is the case for the earlier property of *being-red-or-blue*. This very limited kind of case is the real issue for the strong proportionalist. I argue, though, that it's not such an issue. It's merely an accident of language in these cases that we can't refer to the disjunction with a single term.[19] One possible explanation of this calls upon the utility of an economical language – one which doesn't multiply terms unnecessarily. We could have introduced a single term for things that are red or blue, which would then allow us to pick out the cause in the example case with a single term. But the utility of such a term fails to justify its introduction. The example case is a weird one, and for cases like this we can simply employ the 'or' operator, albeit sacrificing whatever utility is produced by being able to identify causal relata with single terms.

§5.7 Conclusion

I have showed how on my proposed definition of proportional causation, backed by my account of how causal models represent, the strong principle of proportionality can respond to both Franklin-Hall's objection and the problem of generic causes. While much of the problem of disjunctive causes similarly dissolves, the strong proportionalist does need to concede the kind of disjunctive cause where the disjuncts are mutually exclusive properties. Sometimes we have a single term for such a disjunct, but sometimes not. I've argued that it's merely an accident of language when not.

[19] Thanks to David Papineau, in discussion, for this point.

CHAPTER 6

The Importation Problem for Interventionist Semantics

> Contrasted with the [Lewis-Stalnaker] 'possible worlds' account of counterfactuals, this 'structural' model enjoys the advantages of representational economy, algorithmic simplicity, and conceptual clarity.
>
> (Pearl, 2013, p. 977)

§6.0 Abstract Structural equation models lend themselves to a semantics of counterfactuals. Call this an *interventionist semantics* of counterfactuals. It is thought that such a semantics improves on a traditional similarity semantics in that it straightforwardly incorporates causal structure and avoids talk of a similarity relation between possible worlds (Pearl, 2000, 2013; Starr, 2019). This chapter shows, however, that a structural equation analysis of counterfactuals is vulnerable to the same fundamental problem as is a similarity analysis – the problem of identifying what information is relevant to a counterfactual evaluation (Goodman, 1955; Priest, 2018). I argue that where similarity semantics relies on an unarticulated notion of similarity, an interventionist semantics relies on an unarticulated notion of aptness.

§6.1 Introduction

Structural equation models have recently been the source of a promising development in providing a semantics for counterfactuals.[1] Call any such semantics an *interventionist semantics*.[2] Prior to this development, the most popular semantics has been a Lewis-Stalnaker style semantics that provides truth-conditions for counterfactuals in terms of similarity across possible worlds (Lewis, 1973a, 1973b, 1979, 1986; Stalnaker, 1968). Call such a semantics a *similarity semantics*. However, proponents of similarity semantics have had difficulty providing a satisfactory account of how similarity is measured in such a way that captures our intuitions about counterfactuals.

In particular, one family of counterexamples seems to demand the incorporation of causal information into the semantics (Barker, 1999; Edgington, 2004; Schaffer, 2004). (I will discuss these counterexamples in §6.2.a.) Given this, structural equation models, which explicitly encode causal information, suggest themselves as a natural tool for providing an improved semantics.[3]

[1] To be clear about what should be obvious, a semantics for counterfactuals in terms of structural equation models must take a position on the debate about how to interpret the equations of a SEM. In particular, equations *must* be taken to represent type-level causal dependencies, *not* complex counterfactuals. I will therefore break my neutrality on this debate and, for this chapter at least, assume that equations represent type-level causal dependencies.

[2] For various semantics that incorporate structural equation models, see (Briggs, 2012; Ciardelli et al., 2018; Galles & Pearl, 1998; J. Halpern, 2000; Hiddleston, 2005a; Huber, 2013; Kaufmann, 2013; Pearl, 2000; Santorio, 2014, 2019; Schulz, 2011; Woodward, 2003).

[3] Structural equation models can be used either as a way to precisify the similarity relation within the framework of a similarity semantics or as providing a new framework entirely. To simplify the exposition, I will call *any* semantics that incorporates these models an 'interventionist semantics', regardless of background framework – whether it's provided by similarity semantics, premise semantics, or the models themselves. The argument in this chapter applies to any such semantics.

Interventionist semantics has some limitations. In its most developed state (Briggs, 2012), it still isn't able to evaluate iterated counterfactuals with a counterfactual in the antecedent. It also cannot evaluate counterfactuals where the consequent is non-causally dependent on the antecedent, such as "had the table been made of ice, then the legs would have been made of ice."[4] Even so, it is thought that such a semantics at least improves on traditional similarity semantics in that it straightforwardly incorporates causal structure and avoids reliance on the notoriously messy notion of a similarity relation between possible worlds (Pearl, 2000, 2013; Starr, 2019).

Let us concede that an interventionist semantics avoids reliance on a similarity relation between possible worlds. I argue that it nevertheless suffers from the same fundamental problem as a similarity semantics. This is the problem of identifying what information is relevant to a counterfactual evaluation. Originally recognized as the problem of cotenability within Goodman's framework (1955), the more general problem has recently been coined *the importation problem* by Priest (2018). I adopt this label for its generality.

In what follows, I first provide an overview of similarity semantics and how it is susceptible to the aforementioned family of counterexamples, and then an overview of interventionist semantics and how it is meant to resolve them. I then discuss the importation problem, and how each semantics respectively responds. I claim that the best response for an interventionist semanticist relies on an opaque notion of *aptness* in precisely the same way

[4] Counterfactuals like these cannot be evaluated using causal models because the antecedent and consequent cannot both be represented in the same model. They cannot be represented by the same variable due to violating exclusivity, nor by distinct variables due to violating distinctness.

that a similarity semantics relies on an opaque notion of *similarity*. The difficulties associated with articulating aptness are on a par with those associated with providing a similarity measure. I conclude with a series of cases that serve to illustrate these difficulties, indicating where they might be resolved with recourse to my earlier account of aptness.

§6.2 Counterfactual Semantics and the Importation Problem

6.2.a Similarity Semantics

The term 'counterfactual' is perhaps most naturally understood as referring to a conditional statement with a contrary-to-fact antecedent. However, I'll follow the literature in employing the term 'counterfactual' to cover any subjunctive conditional. An exemplar counterfactual is a conditional of the form 'If it had been the case that *A*, then it would be the case that *C*.'[5]

Previously suspect, counterfactuals gained legitimacy with the development of a semantics in the later 20th century by David Lewis and Robert Stalnaker, working independently (Lewis, 1973a, 1973b, 1979, 1986; Stalnaker, 1968). Now known alternatively as Lewis-Stalnaker style semantics, possible worlds semantics, or similarity semantics, it says, roughly, that a counterfactual $A \Box\!\!\rightarrow C$ is true just in case for any world where *A* is true and *C* is false, there is a world more similar to the actual world where *A* and *C* are both true. There

[5] Some challenge whether the best semantics of counterfactuals gives them a ternary structure – an operator connecting two separate propositions. However, both similarity and interventionist semantics treat counterfactuals in this way. So, I'll set aside the alternatives. See (Bennett, 2003, pp. 6–7) for discussion and references.

is then the question of how to measure similarity. It is well-known that a common sense notion of overall similarity fails. This is helpfully illustrated by a counterexample presented in Fine (1975): Assume that the nuclear button is in good working order. Even though Nixon doesn't in fact press this button, the following counterfactual seems true:

> ***A***: Had Nixon pressed the nuclear button, there would have been a nuclear war.

However, a world in which Nixon pressed the button and it results in a nuclear war – call this w_1 – is overall less similar to the actual world than one in which Nixon presses the button but the button spontaneously fails and so there is no nuclear war – call this w_2. After all, w_1 is radically dissimilar from the actual world onward from the point of time at which the button is pressed. W_2, on the other hand, diverges from the actual world only in the small miracle of the button failing, and otherwise almost perfectly matches the actual world from this point of time onward.

In response, Lewis (1979) puts forward the following guide for how to measure similarity, the precise details of which are permitted to vary with context:

(1) It is of the first importance to avoid big, widespread, diverse violations of law.
(2) It is of the second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails.
(3) It is of the third importance to avoid even small, localized, simple violations of law.

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (1979, p. 472)

This produces the result that w_1 is in fact more similar to the actual world than w_2 in the following way. The fact that the button fails in w_2 has relatively trivial but otherwise real and wide-reaching consequences. These consequences mean that there is not perfect match but merely *approximate* match between w_2 and the actual world from the time when the button fails onward. But according to the new guide, merely approximate match is of the least importance, if of any at all. These many consequences could be suppressed through the operation of many small miracles, but this would constitute a significant cost in similarity.[6]

Unfortunately, this guide falls short in the face of even further counterexamples.[7] Take the following case as an example: Assume that the law governing coin tosses is indeterministic. David places his bet on the toss – heads. Dorothy tosses a coin and it lands tails. David has lost the bet. Dorothy says:

B: Had you bet tails, you would have won.

Intuitively, ***B*** is true. Had David bet tails, he would have won. But similarity semantics notoriously struggles to capture this intuition. Take first the set of worlds which perfectly match the actual world in terms of both laws and matters of fact up until just before the bet,

[6] Lewis (1986) alters his account again to deal with counterexamples involving indeterminism (Slote, 1978), but the altered account is susceptible to even further counterexamples (Elga, 2001; Hawthorne, 2005; Wasserman, 2006). I'll keep to the simpler account as further iterations don't affect the dialectic in which I'm interested.

[7] See (Edgington, 2004; Schaffer, 2004) for thorough surveys of these kinds of counterexamples to a similarity semantics.

when a small miracle occurs that makes it the case that David bets tails instead of heads. Then, those worlds that unfold from here in perfect accord with the laws of the actual world are more similar, while those with any law violations are less similar. Accordance with the laws requires that the toss is governed by the same indeterministic law as it actually is. How the toss *lands* in a world is merely a matter of approximate match and is therefore "of little or no importance (Lewis, 1979, p. 472)." Thus, worlds in which the coin lands heads will be equally similar to the actual world as those in which it lands tails. Consider, then, two worlds – w_3, in which the toss lands tails, and w_4, in which it lands heads. ***B*** is true in w_3 but false in w_4. So, similarity semantics renders ***B*** false.

To see what's gone wrong, consider that our intuitive evaluation of ***B*** seems to hold fixed the fact of the coin landing tails. But why do we do this? Consider how we evaluate the alternative counterfactual:

> ***C***: Had David bet tails, snatched the coin out of mid-air, and flipped it, then he would have won.

C is clearly false. Had David snatched the coin out of mid-air and flipped it, it may have come down tails but may just as easily have come down heads. In our evaluation of ***C***, we don't hold fixed the fact of the coin landing tails. The obvious explanation of this is that the antecedent in ***C*** explicitly disrupts the causal history of the coin landing tails. Thus disrupted, we no longer hold fixed what is causally downstream – namely, the outcome of the toss. The antecedent in ***B***, on the other hand, leaves the causal history of the toss alone. As a fact causally independent of the antecedent, the outcome of the toss is held fixed.

Arguably, then, similarity semantics is susceptible to these counterexamples because it is insensitive to causal structure.[8] In general, it seems that our commonsense evaluation of a counterfactual counts variation only in what is causally *independent* of the antecedent as making for dissimilarity, but not in what is dependent. In the actual world, how Dorothy's toss lands is causally independent of David's bet. So, the most similar worlds will preserve the outcome of the toss. As a result of the difference in toss outcome, w_4 is ruled as more dissimilar to the actual world than w_3. ***B*** comes out true when we attend to causal structure.

A similar moral seems to apply to the Nixon case. The radical dissimilarity of a possible world such as w_1 in which nuclear war breaks out counts for nothing, since these dissimilarities are causally dependent on the antecedent – on Nixon pressing the button. A world such as w_2, on the other hand, requires a small miracle to bring about the failing of the button. Since the failing of the button is causally independent of Nixon's pressing it, this makes w_2 more dissimilar than w_1 to the actual world. ***A*** comes out true when we attend to causal structure.

Let's suppose this is correct – that causal structure plays a role in our evaluation of counterfactuals. One could respond by incorporating causal structure into a similarity semantics. Indeed, Schaffer (2004) provides one version of this possibility. But this still

[8] For arguments that incorporating causal information into the counterfactual evaluation will resolve such counterexamples, see (Barker, 1999; Edgington, 2004; Hiddleston, 2005a; Kvart, 1986; Schaffer, 2004).

leaves us relying on an unsatisfyingly vague measure of similarity. Interventionist semantics, however, does away with similarity and already straightforwardly incorporates causal structure. As a result, interventionist semantics has emerged as a promising alternative to similarity semantics.[9]

6.2.b Interventionist Semantics

An interventionist semantics can be motivated by the fact that causal information is thought to be well modelled by structural equation models (SEMs). Invoking the SEM framework, an interventionist semantics says that the truth-conditions for counterfactuals are based on what comes out true in a SEM when the antecedent is set by intervention. More carefully, a counterfactual, A □→ C, is true just in case there is an apt model-interpretation pair, representing A as $X = x$ and C as $Y = y$, according to which $Y = y$ when $X = x$ is set by intervention, and A □→ C is false otherwise. Leave aside for now the question of how to define 'apt' for this purpose, which I'll pick up again in the next section. Suffice it here to recognize that an interventionist semantics will need some notion of aptness or other.

Let's see how this semantics evaluates our earlier examples. Take the betting example first. The scenario is naturally modelled with the following SEM and interpretation.

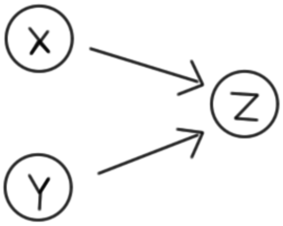
[9] Note also that incorporating causal structure into a similarity semantics would mean giving up on the reductive project of reducing causal dependence to counterfactual dependence. And while an interventionist semantics renders equally impossible a reductive project in this direction, it would allow for a partial reduction in the other direction – namely, reducing at least some counterfactual dependencies to causal.

Figure 17.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{X, Y\}$
$\boldsymbol{V} = \{Z\}$
$\boldsymbol{R} = f\,(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 1$
(EQ2) $Y = 0$

$\boldsymbol{\mathcal{L}}$ = (EQ3) $Z := \begin{cases} 1\ if\ X = Y \\ 0\ if\ X \neq Y \end{cases}$



$$X\ (David) : \begin{cases} 1\ if\ betting\ heads \\ 0\ if\ betting\ tails \end{cases} \qquad Y\ (coin) : \begin{cases} 1\ if\ landing\ heads \\ 0\ if\ landing\ tails \end{cases}$$

$$Z\ (David) : \begin{cases} 1\ if\ wins\ the\ bet \\ 0\ if\ loses\ the\ bet \end{cases}$$

$\boldsymbol{\mathcal{A}}$ assigns *X* the value *1* and *Y* the value *0*. This represents the actual facts of David betting heads and the coin landing tails. Here, setting the antecedent by intervention is analogous to setting the variable *X* to *0*, since '*X = 0*' represents that David bets that the coin will land tails. The value of *Y* is unchanged. *Z*'s value is partially determined by that of *X*, and so its value is updated in response to the intervention on *X*. Since *X* is now *0*, and *Y* has always been *0*, *X* = *Y*. Thus, the value of *Z* is *1*. '*Z = 1*' means that the consequent of ***B*** is true since it represents that David has won the bet. So, interventionist semantics delivers the verdict that ***B*** is true – the desired result.

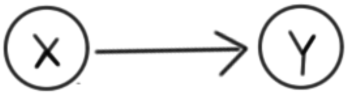
Next, take the Nixon example. The scenario is naturally modelled in the following way.

Figure 18.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U}$ = $\{X\}$
$\boldsymbol{V}$ = $\{Y\}$
$\boldsymbol{R} = f(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $X = 0$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $Y := X$



$$X\ (button) : \begin{cases} 1\ if\ pressed \\ 0\ if\ not\ pressed \end{cases} \qquad Y\ (nuclear\ war) : \begin{cases} 1\ if\ occurs \\ 0\ if\ doesn't\ occur \end{cases}$$

X is assigned the value *0* since Nixon doesn't press the button in the actual situation. Then, in order to evaluate ***A***, we set the antecedent by intervention. So, we set the variable *X* to *1*, since '*X* = *1*' represents that Nixon presses the nuclear button. As a result, *Y* takes the value *1*, which represents that nuclear war occurs. This means that the consequent of ***A*** holds when the antecedent is set by intervention. Interventionist semantics determines ***A*** to be true – the desired result.

Now, these truth-conditions are complicated by the possibility of there being more than one apt model-interpretation pair. A choice must be made about how to handle this. One can endorse a universal principle whereby a counterfactual will be true just in case the consequent is true *in every* apt model-interpretation pair wherein the antecedent is set by intervention, and false otherwise. One could instead endorse an existential principle whereby a counterfactual is true just in case *there is* an apt model-interpretation pair in which the consequent is true. And one could endorse any of myriad principles in between –

a counterfactual is true just in case the consequent is true in the majority of apt model-interpretations, in 60% of apt model-interpretations, in some special subset of them, etc. I'll assume the existential principle, since it is widely adopted in discussions of actual causation.[10]

One might think that by requiring just one model-interpretation pair to determine a counterfactual as true, the existential principle permits a less demanding notion of aptness than the universal one. After all, we won't need our notion of aptness to rule out rogue model-interpretations which deliver the result of false, since they won't undermine an otherwise true verdict. We would, though, on the universal principle. But this thought is short-sighted. Either principle requires the same amount of work of aptness. It's just that the work will be slightly different. A universal principle means that truth is hard to come by while falsity is relatively easy. So, aptness would need to rule out model-interpretation pairs which render false counterfactuals we deem true. The existential principle, on the other hand, means that truth is relatively easy to come by while falsity is difficult. So, aptness would need to rule out model-interpretation pairs which render true counterfactuals we deem false. This last will be the challenge addressed in this paper.

6.2.c The Importation Problem

[10] As discussed in §1.5. Views of actual causation that adopt an existential principle include (Blanchard & Schaffer, 2017; Hitchcock, 2001, 2009, 2018; Weslake, 2015; Woodward, 2003).

An interventionist semantics may seem to be an obvious improvement over a similarity semantics in that it doesn't rely on ambiguous talk of similarity. But while it's true that an interventionist semantics doesn't rely on a notion of similarity, I'll argue that it does rely on an analogous notion – that of *aptness*.

Let us begin by stepping back. A counterfactual links an antecedent with a consequent. On any semantics, the evaluation of a counterfactual will minimally include the truth of the antecedent. But this is yet insufficient to render a verdict. If the antecedent is actually false, then some other aspects of the actual situation will need to be altered to accommodate the truth of the antecedent. Determining which aspects are held fixed alongside the antecedent and which are left behind is arguably *the* fundamental problem for a semantics of counterfactuals. Nelson Goodman explains,

> The first major problem [concerning counterfactuals] is to define relevant conditions: to specify what sentences are meant to be taken in conjunction with an antecedent as a basis for inferring the consequent. (1955, p. 8)

And Dorothy Edgington writes,

> In trying to settle the matter [of whether C would have been true given A], you need to rely on some actual facts, and let other facts go by the board with the supposition that A. What determines what you can hang on to, and what you must give up? (2004, p. 13)

While widely recognized, the general version of this problem seems to have only recently been given a name. Priest (2018) coins it *the importation problem*. In terms of information, the question is: what information should be carried over – or imported – from the actual situation into the situation(s) relevant to counterfactual evaluation? Let's call this the *importation question*.

A similarity semantics answers this question by its similarity metric. In general, it says that all laws and matters of fact necessary for achieving *relevant similarity* to the actual situation should be imported, and nothing else. By still relying on a notion of relevance, this is as yet an incomplete answer. Lewis attempts to fill this in with his levels of importance of similarity: it is of the utmost importance to import major laws, of the next importance to import perfectly matching spacetime regions, of the third importance to import minor or local laws, and finally of the least or even no importance to import approximately matching spacetime regions. While some help as a guide, the coin example illustrates one way in which this answer to the importation question falls short.[11]

A naïve thought from the interventionist semantics camp may be that an interventionist semantics is not susceptible to the importation problem. Counterfactuals are straightforwardly evaluated relative to models. But the importation problem arises for interventionist semantics in the question of *which* model(s) on *which* interpretation(s). Recall that the semantics indicates a domain of *apt* model-interpretation pairs over which

[11] For further examples, again see (Edgington, 2004; Elga, 2001; Hawthorne, 2005; Schaffer, 2004; Slote, 1978; Wasserman, 2006).

the truth-conditions existentially quantify: a counterfactual, $A \Box\!\!\rightarrow C$, is true just in case there is an apt model-interpretation pair, representing A as $X = x$ and C as $Y = y$, according to which $Y = y$ when $X = x$ is set by intervention, and $A \Box\!\!\rightarrow C$ is false otherwise. But it does not yet say which ones are apt. This problem isn't immediately apparent in the literature on interventionist semantics, because counterfactuals are either evaluated relative to a 'natural' model – without any principled account of why that model is so qualified, or else systematically stated in terms that already presuppose a model – such as "$\textit{Flame} = 1 \Box\!\!\rightarrow \textit{Meat cooked} = 1$ (Hitchcock, 2018, p. 17)." But natural language counterfactuals are not already formulated in terms that presuppose a model, and so must be translated into the terms of some model or other via an interpretation. An interventionist semantics needs to provide a translation guide that adjudicates first on what makes an interpretation admissible, and then on whether the model being interpreted is appropriate or not – is 'apt' or not. The importation question for an interventionist semantics is: which aspects of the actual situation should be represented by the model? As I'll argue in the next section, answering this question is far from straightforward.

§6.3 The Problem of Aptness

The literature has gone some way towards a full articulation of aptness, at least for the purposes of defining causation.[12] However, more work is required to provide a satisfactory account, and the nature of this work is recognizably messy in precisely the same way as is

[12] See (Blanchard & Schaffer, 2017; N. Hall, 2007; J. Halpern & Hitchcock, 2010; J. Y. Halpern, 2016b, 2016a; Hitchcock, 2001, 2007a, 2012; Menzies, 2017; Woodward, 2016)

work on similarity. Hitchcock writes that "[w]hat constitutes an appropriate model is a tricky affair, more a matter of art than science." (2007a, p. 503) Woodward concedes that aptness may indeed be an art, yet he explores the "opposite" position. This he describes as the position "that there are useful (although of course defeasible) heuristics/default rules of some generality that can be used to guide variable choice (2016, p. 1048)," acknowledging that "such heuristics are unlikely to yield, in most cases, a uniquely best choice of variables…(2016, p. 1048)" Finally, after providing several conditions of aptness, Blanchard and Schaffer write,

> We would emphasize that all of these are vague conditions, aspects of the art rather than the science of causal modelling. In no case does one find a mathematically precise account of these conditions within the terms of the structural equations framework. Rather these are extra-mathematical conditions on the relation between the mathematics and the reality it would represent. Do not expect more. (2017, p. 183)

In the remainder of this paper, I will illustrate the difficulty with articulating aptness for evaluating counterfactuals by presenting three cases and some problems that arise in constructing models for them. Where applicable, I will show how a response to these problems can be given from my proposed account of aptness.

6.3.a Real, Relevant, and Essential Causal Links

In the first case, two delinquent children, Suzy and Billy, walk past an abandoned house. Suzy picks up the only brick lying around and throws it at a window. The brick hits the window and the window shatters. Consider the following counterfactual:

W: If Suzy had not thrown the brick, then the window would not have shattered.

Let's assume that in this case, had Suzy not thrown the brick then Billy would have. Suzy's throwing the brick satisfies Billy's delinquent intentions. But had Suzy not thrown the brick, then Billy would have thrown it – driven by these very delinquent intentions. And let's adopt the simplifying assumption that either child will only throw with perfect accuracy and with sufficient force so as to overcome the force holding the window together given the mass of the thrown brick.[13] Given all of this, ***W*** is false. Had Suzy not thrown the brick, then Billy would have thrown it and the window would still have shattered.

However, there are infinitely many model-interpretation pairs with which we could evaluate ***W***. After all, for any model we might construct and interpret, we could always simply add another variable and produce a new, distinct pair. Of course, adding irrelevant variables shouldn't affect the counterfactual evaluation. In itself, a plethora of possible interpreted models need not be problematic – but it is when one of them delivers a true verdict for a counterfactual that we want to come out false, such as ***W***.[14] Aptness is needed to rule these out. Let's see how that goes.

[13] Such simplifying assumptions are pervasive in the causal model literature and are generally benign.

[14] This of course assumes the existential principle. Given the universal principle, the problem would instead arise when a model delivers a false verdict for a true counterfactual.

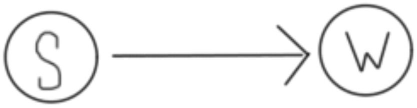
In order to evaluate $\boldsymbol{W}$, I'll start with a model that has two variables, S and W, interpreted with S representing Suzy's throwing her brick or not, and W representing the window shattering or not. The exogenous variable, S, is assigned the value *1*, to represent the actual fact of Suzy throwing the brick. Call this $\mathcal{M}_5$.

Figure 19.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{S\}$
$\boldsymbol{V} = \{W\}$
$\boldsymbol{R} = f\,(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $S = 0$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $W := S$



$\boldsymbol{\mathcal{M}_5}$

$$\mathcal{I}(\mathcal{M}_5):\qquad S\,(Suzy): \begin{cases} 1\ if\ throws \\ 0\ if\ doesn't\ throw \end{cases}\qquad W\,(window): \begin{cases} 1\ if\ shattered \\ 0\ if\ not\ shattered \end{cases}$$

The antecedent of $\boldsymbol{W}$ is then set by intervention. So, S is set to *0* to represent Suzy refraining from throwing. This results in W also taking the value *0*, which represents the window not shattering. Thus, $\mathcal{M}_5$ on $\mathcal{I}(\mathcal{M}_5)$ delivers the verdict that $\boldsymbol{W}$ is true. Since we want $\boldsymbol{W}$ to come out false, $< \mathcal{M}_5, \mathcal{I}(\mathcal{M}_5) >$ must be inapt for some reason. Several reasons spring to mind, but it is another question whether they can be codified in a principle of aptness.

The first reason might be that $< \mathcal{M}_5, \mathcal{I}(\mathcal{M}_5) >$ seems to say falsely that the causal link between the window shattering and Suzy's throw is such that whether or not the window shatters is fully causally determined by whether or not Suzy throws. Turning this into a principle of aptness is straightforward: the equations of an apt interpreted model should be *true*. Since an interventionist semanticist interprets the equations as representing type-level causal dependencies, or causal *links*, this means that an apt interpreted model should represent only real causal links – that is, causal links that actually exist.[15] The argument would then go that this requirement renders $< \mathcal{M}_5, \mathcal{I}(\mathcal{M}_5) >$ inapt for this situation, since *W := S* does not represent a real causal link.

But is it really the case that *W := S* fails to represent a real causal link? Suzy's throw is indeed of the type to cause a window shattering. Furthermore, it is of the precise type to cause the precise type of window shattering. Whether *W = 1* is taken to represent the general fact of the window shattering or the finely detailed shattering that actually happens, *W := S* represents a real causal link.

While independently plausible, the aptness principle requiring that the equations of an apt interpreted model be true doesn't do enough to help in this case. One might instead attempt to explain what's wrong with $< \mathcal{M}_5, \mathcal{I}(\mathcal{M}_5) >$ by pointing out that *W := S* isn't the only *relevant* dependency in the area for evaluating ***W***. Billy's throw is *also* of the type to cause a window shattering. Perhaps we need something like a principle that requires that *all relevant* dependencies are represented. However, while also plausible, this is less illuminating than we need. The notion of *relevance* is as ambiguous as that of similarity.

[15] This principle has an analogue for those not in the business of an interventionist semantics – those who take the equations to represent complex counterfactuals rather than causal links. The analogous principle of aptness is that the counterfactuals encoded by a model be true. This is endorsed by Woodward (2016, p. 1055), and corresponds to the first necessary condition on aptness for actual causation put forward by Hitchcock (2001, p. 287), and picked up by Blanchard and Schaffer (2017, p. 182).

One might be tempted by a principle of aptness that simply requires that all parts of the situation introduced in the description be represented by the model. This would ensure that Billy be represented. Indeed, it would ensure that everything "relevant" be represented without seeming to invoke that notion. Such a principle may lead to irrelevant aspects being represented, of course, but perhaps this would not lead to any real problems. However, this line of reasoning is short-sighted. Such a principle would work only on the assumption that a given description of a situation captures everything relevant about that situation. While this may be the case in practice, it is far from guaranteed in principle. As competent language speakers, we do tend to capture all relevant aspects when describing a situation. But this feature is not inherent in the nature of a description. The principle would need to specify that the type of description involved is one that captures all relevant aspects. Looks like we would need to invoke relevance after all.[16]

Of course, a straightforward principle that would do the trick is one that requires that enough variables be included so that the counterfactual comes out right. But this would make the truth-value of the counterfactuals determine the relevant notion of aptness, when what we wanted was for the models to provide an independent grounding for the counterfactuals. Such a principle would undermine the project.

Another way to put the problem with $<\mathcal{M}_5, \mathcal{I}(\mathcal{M}_5)>$ is that it seems to leave some crucial information out – namely, the presence of Billy and his inevitable response to Suzy's failure to throw. This suggests an extant principle of aptness in the literature might help – the one

[16] Thanks to an anonymous referee for this suggestion.

that requires that enough variables be included so as to capture the *essential structure* of the situation being modelled (Blanchard & Schaffer, 2017, p. 183; Hiddleston, 2005a, pp. 648–649; Hitchcock, 2007a, p. 503). But the notion of essential structure is not much of an improvement on that of relevant dependencies. We know pre-theoretically that Billy factors into the essential structure of this situation, but what underlying principle can we give to justify this? Without a way of delineating what counts as *essential* in any given situation, this principle also remains less illuminating than we'd like. In order to improve on similarity, aptness principles need to be cleaner - clearer and more objective or more principled in some way. Reliance on notions like relevance or essential does not yet achieve this.

This last way of putting the problem suggests that the aptness principle proposed in Chapter 3 would help in this case. The principle of Explicit Partial Mediation requires that any feature be explicitly represented by the model if, were it to be represented, it would be represented by a partially mediating variable. Billy's throw, were it to be represented in $\mathcal{M}_5$, would be represented by a partially mediating variable. Explicit Partial Mediation (EPM) delivers the result that it should therefore be explicitly included in any apt interpreted model. Here, at last, is an objective aptness principle that rules $\mathcal{M}_5$ inapt – due to violating EPM.

6.3.b Introducing Distinctness

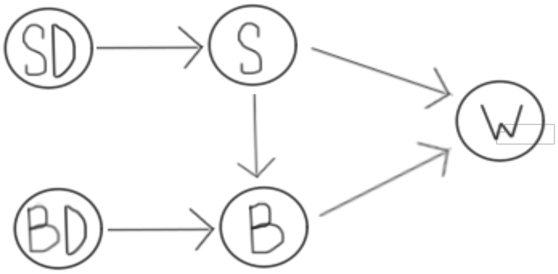
The initial problem can be solved by invoking Explicit Partial Mediation. However, in constructing a model that respects EPM, we run into a dilemma involving whether or not to require *distinctness*. To see this, let's first construct what strikes me as a natural model for evaluating $\boldsymbol{W}$ relative to the target situation. The new model has three new variables in addition to *S* and *W*: *SD*, *BD*, and *B*. *SD* represents Suzy's having a delinquent intention or not, *BD* represents Billy's having a delinquent intention or not, and *B* represents Billy throwing the brick or not. The exogenous variables are now *SD* and *BD*, and they are each assigned the value *1*, to represent the actual fact of Suzy and Billy having delinquent intentions. Call this next model $\mathcal{M}_6$, and the corresponding interpretation $\mathcal{I}(\mathcal{M}_6)$.

Figure 20.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U}$ = $\{SD, BD\}$
$\boldsymbol{V}$ = $\{S, B, W\}$
$\boldsymbol{R} = f\,(X_i) = \{1, 0\}$

$\boldsymbol{\mathcal{A}}$ = (EQ1) $SD = 1$
(EQ2) $BD = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ3) $S := SD$
(EQ4) $B := \min\,(BD, (1 - S))$
(EQ5) $W := \max\,(S, B)$



$\boldsymbol{\mathcal{M}_6}$

$\mathcal{I}(\mathcal{M}_6)$: $SD\ (Suzy) : \begin{cases} 1\ if\ has\ a\ delinquent\ intention \\ 0\ if\ doesn't\ have\ a\ delinquent\ intention \end{cases}$

$BD\ (Billy) : \begin{cases} 1\ if\ has\ a\ delinquent\ intention \\ 0\ if\ doesn't\ have\ a\ delinquent\ intention \end{cases}$

$S\ (Suzy) : \begin{cases} 1\ if\ throws\ the\ brick \\ 0\ if\ doesn't\ throw \end{cases}$ $B\ (Billy) : \begin{cases} 1\ if\ throws\ the\ brick \\ 0\ if\ doesn't\ throw \end{cases}$

$W\ (window) : \begin{cases} 1\ if\ shattered \\ 0\ if\ not\ shattered \end{cases}$

Again, we evaluate $\boldsymbol{W}$ by setting the antecedent by intervention. So, variable S is set to *0*. This results in variable B taking the value *1*, and W taking the value *1*, which represents that the window shatters. Thus, $\mathcal{M}_6$ on this interpretation evaluates $\boldsymbol{W}$ as false, as desired. So, $\boldsymbol{W}$ is *possibly* false, so long as it doesn't come out true in any other apt interpreted model.

Unfortunately, while providing a correct and intuitive evaluation of $\boldsymbol{S}$, $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ violates *distinctness*, which requires that distinct variables represent genuinely distinct things. According to $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$, if S takes value *1*, then B *cannot* take value *1* – it *must* take value *0*. $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ represents this with a causal link, but it isn't causal. Since there's only one brick in the above scenario, Suzy's throw metaphysically entails that Billy doesn't throw, and vice versa. The brick cannot be in two places at once. A commitment to distinctness requires that Suzy's throwing and Billy's throwing be represented by the same variable, since these are mutually exclusive possibilities. Since $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ fails to do this, it fails to respect distinctness, and is therefore inapt. Before constructing a model that does respect distinctness, I will first explore what happens if we simply reject it.

6.3.c Violating Distinctness

A possible response here would be to deny the need for distinctness. After all, distinctness comes from attempts to use causal models to define *actual causation* (as in Chapter 2), which is a different task than the one in which we're now engaged – namely, using them to provide a counterfactual semantics. Different tasks may call for different representational principles.[17] By rejecting distinctness, we would dissolve the reason for ruling $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ inapt.

Unfortunately, distinctness seems to still be needed. The following case illustrates the type of problem we run into by rejecting distinctness, even in the context of evaluating for counterfactual dependency rather than actual causation.

Two students, Alice and Betty, are placing their votes for two candidates: Xavier and Yancy.[18] Alice has a green ballot and Betty has a red one, so whomever Alice votes for will receive a green vote and whomever Betty votes for will receive a red one. Although they vote entirely independently, Alice and Betty happen to both vote for Xavier, so Xavier receives both a green vote and a red one.[19] Yancy receives no votes. Consider the following counterfactual:

> ***V***: If Betty had voted for Yancy, then Yancy would have received a green vote.

Betty's ballot is red, not green. Intuitively, then, ***V*** is false. Had Betty voted for Yancy, then Yancy would have gotten a *red* vote, not a green one. But say we model this in the following way, which I'll call $\mathcal{M}_7$. $\mathcal{M}_7$ has four variables. $\mathcal{I}(\mathcal{M}_7)$ interprets them in the following way. B represents Betty's voting for Xavier or for Yancy, S represents Alice and Betty voting for the same candidate or for different ones, G represents Xavier receiving the green vote or Yancy

[17] Blanchard and Schaffer make this point with respect to different notions of causation (2017, p. 181).

[18] This case is adapted from (J. Y. Halpern, 2016b) (who in turn adapts it from (N. Hall, 2007)).

[19] The independence of Alice's and Betty's vote ensures that the exogenous variables are relevantly independent – that is, the types instantiated by S and by B are probabilistically independent in the way required by Pearl (2000, p. 27). S is only probabilistically dependent on B *conditional on* how Alice votes.

receiving it, and R represents Xavier receiving the red vote or Yancy receiving it. The exogenous variables, B and S, are each assigned the value *1*, representing the actual facts of Betty voting for Xavier and of Alice and Betty voting the same.

Figure 21.

$\mathcal{S}$ = **U** = $\{S, B\}$

V = $\{G, R\}$

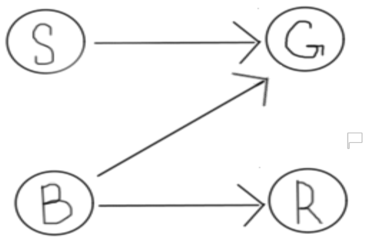
R = $f(X_i) = \{1, 0\}$

$\mathcal{A}$ = (EQ1) $S = 1$

(EQ2) $B = 1$

$\mathcal{L}$ = (EQ3) $G := \begin{cases} 1\ if\ S = B \\ 0\ if\ S\ \neq B \end{cases}$

(EQ4) $R := B$



$\mathcal{M}_7$

$\mathcal{I}(\mathcal{M}_7)$: *B (Betty)* = $\begin{cases} 1\ if\ votes\ Xavier \\ 0\ if\ votes\ Yancy \end{cases}$ *S (Alice and Betty)* = $\begin{cases} 1\ if\ vote\ the\ same \\ 0\ if\ vote\ differently \end{cases}$

R (red vote) = $\begin{cases} 1\ if\ received\ by\ Xavier \\ 0\ if\ received\ by\ Yancy \end{cases}$ *G (green vote)* = $\begin{cases} 1\ if\ received\ by\ Xavier \\ 0\ if\ received\ by\ Yancy \end{cases}$

Intervening on $< \mathcal{M}_7, \mathcal{I}(\mathcal{M}_7) >$ to set the antecedent would set B to *0*. This change doesn't affect S, since it gets its value exogenously. Since $B = 0$ and $S = 1$, $S \neq B$, and so G takes the value *0.* Since '$G = 0$' represents that Yancy receives the green vote, the consequent holds. $< \mathcal{M}_7, \mathcal{I}(\mathcal{M}_7) >$ delivers the verdict that ***V*** is true – not the desired result.

The reason for our trouble here is precisely that S and B are metaphysically intertwined and so do not satisfy distinctness. Whether Alice and Betty vote the same or different is a function of how Alice votes and how Betty votes. Crucially, though, it is a *logical* function of their votes, not a causal function. By stipulation, Alice and Betty's votes are causally independent of each other. This choice of metaphysically intertwined variables S and B leads us to evaluate the counterfactual in a strange way. It has us hold fixed the fact that Alice and Betty vote the same, even if Betty were to vote differently than she actually does, despite there being no causal dependency in the world that justifies this. Although Halpern (2016b) bites the bullet here, this seems highly unattractive. Better to rule $< \mathcal{M}_7, \mathcal{I}(\mathcal{M}_7) >$ as inapt due to its violation of distinctness.

So, violating distinctness leads to trouble. Better to require it. Unfortunately, requiring distinctness leads to additional trouble – as I'll demonstrate in the next section.

6.3.d Requiring Distinctness

Refer back to the example of Suzy and Billy and our consideration of $\boldsymbol{W}$:

> $\boldsymbol{W}$: If Suzy had not thrown the brick, then the window would not have shattered.

Again, $\boldsymbol{W}$ is false given previous assumptions. Had Suzy not thrown her brick, then Billy would have thrown, prompted by his unsatisfied delinquent intentions, and the window would have shattered. The trouble here is not simply how to rule out models that give us the wrong results, but how to evaluate this counterfactual at all. Before the discussion in §6.3.c, we might have thought a model like $\mathcal{M}_6$ on $\mathcal{I}(\mathcal{M}_6)$ – which represents Suzy's throw and Billy's throw with separate variables – would be fine to evaluate $\boldsymbol{W}$. According to $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$, $\boldsymbol{W}$ is false. Assuming no other apt interpreted model renders it true, then, $\boldsymbol{W}$ is false tout court, as desired. But $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ violates distinctness. Due to there being just one brick, Suzy's and Billy's throw are metaphysically intertwined. Respecting distinctness, then, their respective throws need to be represented by alternative values of the same variable.

Let's look at a model that does just this. Call it $\mathcal{M}_8$. $\mathcal{M}_8$ has two variables, T and W. $\mathcal{I}(\mathcal{M}_8)$ interprets T as representing the three distinct possibilities of Suzy throwing, Billy throwing, or neither child throwing the brick, and W as representing the window shattering, as before. We don't include the fourth logical possibility of both children throwing the brick, since it's not physically possible give. The exogenous variable, T, is assigned the value *1*, representing the actual facts of Suzy throwing.

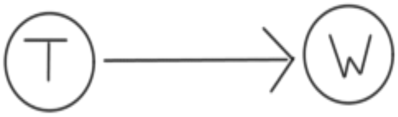
Figure 22.

$\boldsymbol{\mathcal{S}}$ = $\boldsymbol{U} = \{T\}$

$\boldsymbol{V} = \{W\}$

$\boldsymbol{R} = f(T) = \{1, -1, 0\}$

$f(W) = \{1, 0\}$



$\boldsymbol{\mathcal{A}}$ = (EQ1) $T = 1$

$\boldsymbol{\mathcal{L}}$ = (EQ2) $W := |T|$

$\boldsymbol{\mathcal{M}_8}$

$$\mathcal{I}(\mathcal{M}_8): \qquad T\ (brick): \begin{cases} 1\ if\ thrown\ by\ Suzy \\ -1\ if\ thrown\ by\ Billy \\ 0\ if\ not\ thrown \end{cases}$$

$$W\ (window): \begin{cases} 1\ if\ shattered \\ 0\ if\ not\ shattered \end{cases}$$

Despite being accurate of the situation, $\mathcal{M}_8$ on $\mathcal{I}(\mathcal{M}_8)$ cannot yet serve to evaluate $\boldsymbol{W}$. To set the antecedent by intervention, we need to set T to *not-1*. But two options would equally satisfy: T = *-1* and T = *0*. To which value should T be set? This is a choice point for the interventionist semanticist. In order to handle cases where the antecedent can be set by more than one intervention, interventionist semantics needs to be supplemented with one of a range of possible conditions. I'll go through these possible conditions in turn.

Consider first a supervaluationist condition, according to which a counterfactual is true just in case the consequent holds under *every* intervention that sets the antecedent, false just in case the consequent fails to hold under *every* intervention that sets the antecedent, and indeterminate otherwise. Unfortunately, this won't do for our current example. Setting *T = -1* renders the consequent false, but setting *T = 0* renders the consequent true. ***W*** comes out indeterminate, then. But we want ***W*** to come out false. Either the supervaluationist condition is correct but this is an inapt interpreted model (but what makes it inapt?), or else this condition is incorrect.

Alternatively, we could supplement with an existential condition. According to this condition, a counterfactual is true just in case the consequent holds under *at least one* intervention that sets the antecedent, and false otherwise. This also won't do for our current example. Since at least one intervention renders the consequent true (*T = 0*), ***W*** comes out true. This is not the result we're looking for.

What about a universal condition? According to this condition, a counterfactual is true just in case the consequent holds under every intervention that sets the antecedent, and false otherwise. This is the condition Briggs (2012) adopts. It indeed gets the result we're looking for. Since at least one intervention renders the consequent false (*T = -1*), ***W*** is false. But it comes at the cost of also making *false* the following:

> ***W****: If Suzy had not thrown the brick, then the window would have shattered.

Since there is at least one intervention that renders the consequent false (*T = 0*), ***W**** is false. This contradicts the stipulation of the case, that had Suzy not thrown the brick then Billy would have, and so the window would have shattered. Indeed, it's this stipulation which leads us to think that ***W*** is false in the first place. In other words, it's the *truth* of ***W**** that leads us to think that ***W*** is false. The universal condition secures the inference, but undercuts its soundness at the same time.

While distasteful, this result may not be fatal. The semantics is fine so long as we can construct at least one apt interpreted model that renders ***W**** true while keeping ***W*** false. But given we require distinctness, what other model-interpretation pair could there be? A natural thought might be that we could just replace *T* with a new binary variable, *T**, interpreted so as to represent either Suzy throwing or Billy throwing. Call this new model and interpretation $\mathcal{M}_8{}^*$ and $\mathcal{I}(\mathcal{M}_8)^*$. According to $< \mathcal{M}_8{}^*, \mathcal{I}(\mathcal{M}_8)^* >$, ***W**** is true and ***W*** false. However, what justifies the removal of an otherwise physically possible option – the possibility of neither child throwing? If this is simply allowed, then we could have just as easily replaced *T* with *T†*, which also only takes two values – this time, representing Suzy throwing and neither child throwing. According to this new model and interpretation, call them $\mathcal{M}_8{}^{\dagger}$ and $\mathcal{I}(\mathcal{M}_8)^{\dagger}$, ***W**** is false and ***W*** is true. Given that our interventionist truth-conditions existentially quantify over all apt models, we can take the results of $< \mathcal{M}_8{}^{*}, \mathcal{I}(\mathcal{M}_8)^{*} >$ and those of $< \mathcal{M}_8{}^{\dagger}, \mathcal{I}(\mathcal{M}_8)^{\dagger} >$ together, delivering the verdict that both ***W*** and ***W**** are true. This would mean that it is true that had Suzy not thrown the brick, then the window would not have shattered and had Suzy not thrown the brick, then the window would have shattered. Surely an unpalatable result.

We could avoid this result if we could produce a principle of aptness that somehow permits $< \mathcal{M}_8{}^{*}, \mathcal{I}(\mathcal{M}_8)^{*} >$ but excludes $< \mathcal{M}_8{}^{\dagger}, \mathcal{I}(\mathcal{M}_8)^{\dagger} >$ for evaluating ***W*** and ***W****. An obvious principle that would do the trick is one that requires that variables include only *relevant* values, where Billy's throw is considered relevant but neither child throwing is considered irrelevant. But this principle remains unilluminating if we leave the notion of *relevance* unaccounted for, and it seems likely that what underlies it is a set of pragmatic considerations.

It looks, then, like none of the above quantifiers over interventions – a supervaluationist condition, existential condition, nor universal condition – will work in this case. There are other options. We could introduce the condition that a counterfactual is true just in case the consequent holds in *some majority* of the interventions that set the antecedent, or *some important subset* of the interventions, or etc. But it seems that none of these would improve on the three I've already explicated. In particular, any condition that invokes the notion of importance, relevance, or similarity will once again introduce into interventionist semantics the ambiguity it is supposed to avoid. We could instead say that which condition is relevant to a given counterfactual evaluation is somehow determined by context. But whether this could be cashed out in a principled and objective way remains to be seen.

6.3.e The Dilemma

Thus, we have a dilemma. Both requiring distinctness and rejecting distinctness raises problems. Taking the first horn would successfully rule $< \mathcal{M}_7, \mathcal{I}(\mathcal{M}_7) >$ inapt, but it would also rule $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ inapt, thus demanding that we evaluate $\boldsymbol{W}$ with something like $< \mathcal{M}_8, \mathcal{I}(\mathcal{M}_8) >$. This, in turn, would demand somehow identifying the correct quantifier over possible interventions for so doing. Taking the second horn – dismissing distinctness – would legitimize $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ and so permit its use in evaluating $\boldsymbol{W}$, but it would demand finding some other way of explaining what's wrong with $< \mathcal{M}_7, \mathcal{I}(\mathcal{M}_7) >$.

Alternatively, we can respond to this dilemma by challenging one of its presuppositions. Notice that the interventionist truth-conditions are truth-conditions *simpliciter*. But if my argument from Chapter 4 is right, then causal models (in this context) represent type-level causal relations that hold only *relative to a modal profile*. Perhaps the mistaken presupposition is the idea that we can use modally relative causal relations to provide a non-relative counterfactual semantics. Permitting the causal relativity to carry over into the semantics results in a view according to which counterfactuals are true or false only relative to a modal profile. More carefully, a counterfactual, A □→ C, is true relative to a modal profile, θ_i, just in case there is an apt model-interpretation pair that specifies θ_i, represents A as $X = x$ and C as $Y = y$, and according to which $Y = y$ when $X = x$ is set by intervention. A □→ C is false relative to the modal profile, θ_i, otherwise – that is, if there is no such model.

On this view, the dilemma dissolves. We can require distinctness and explain away the dilemma in the following way. First, $< \mathcal{M}_7, \mathcal{I}(\mathcal{M}_7) >$ is inapt since it violates distinctness relative to any modal profile. Next, $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ only violates distinctness relative to a modal profile that holds fixed the fact of there being only one rock. Relative to a modal profile that relaxes this fact, distinctness is satisfied. So, $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ is apt when it specifies the more permissive modal profile, and delivers the verdict that $\boldsymbol{W}$ is false relative to this more permissive modal profile (assuming no other model renders it true relative to this modal profile). Arguably, $< \mathcal{M}_6, \mathcal{I}(\mathcal{M}_6) >$ strikes us as the natural model because this more permissive modal profile is more normal – it is one we would find ourselves in more often.

Unsurprisingly, this is my preferred response. However, although relativity to modal profile dissolves the dilemma, it does not address the deeper issue that launched this discussion. On this view, for any given evaluation of a counterfactual, a modal profile must be selected for. But what determines/justifies the selection of modal profile? In this context, a selection between modal profiles is a selection between different sets of type-level causal relations, each of which is instantiated in the target situation. Notice, then, that this question is simply a more precise articulation of the importation question for an interventionist semantics: which of the many possible sets of causal relations should be imported into the world of evaluation? Thus, while the key representational question for an interventionist semantics may be clarified by a causal relativist view of counterfactuals, it is not resolved.

Each of the above responses to the dilemma that I've laid out calls for further progress down the road of articulating aptness. It has not been my intention to resolve that issue here, but merely to demonstrate that progress is needed and will not be easily won.

§6.4 Conclusion

In sum, let me collect in one place the many possible principles of aptness uncovered by this chapter's inquiry (listed chronologically). A model-interpretation pair is apt only if…

(i) Equations are true – only *real* causal links are represented.

(ii) All *relevant* causal links are represented.

(iii) Enough variables are included so as to capture the *essential structure* of the situation being modelled.

(iv) Anything that, were it represented in the model, would be represented by a variable that partially mediates between two existing variables are included in the model. (*Explicit Partial Mediation*)

(v) Distinct variables represent genuinely distinct things. If two things are not distinct, then they are represented by the same variable.

(vi) Variables include only *relevant* values.

Even assuming each of these conditions is plausible on its own, the list is far from satisfactory. (ii), (iii), and (vi) are unilluminating in precisely the same way that talk of similarity is unilluminating and the dilemma as to whether or not to endorse (v) remains.

In his seminal work, Judea Pearl writes, "[A similarity] semantics still leaves questions of representation unsettled…. Such difficulties do not enter the structural account" (2000, p. 239). Unfortunately, this is overly optimistic. I have shown that there are indeed questions of representation left open. And they are not straightforwardly answered. This mess of unresolved questions of aptness is equal to the mess of similarity for a similarity semantics. Not only is more work required to articulate the notion of aptness, but it seems clear that no formal articulation of it is possible. So much for causal models providing a more determinate semantics of counterfactuals.

One thing this upshot may recommend is to reject causal models as a promising means of providing truth-conditions for counterfactuals. After all, the same thing could be accomplished by simply incorporating causal structure into a similarity semantics. But this discussion has been meant to illustrate the *difficulty* of articulating aptness, not the impossibility of doing so. Thus, some may instead take this as a call to action – a challenge to articulate a better structural equation analysis of counterfactuals. Regardless, interventionist semantics does not yet have the upper hand over similarity semantics.

CHAPTER 7

Conclusion: Recap and Suggestions for Future Research

To conclude, I'll briefly take stock of what's been argued so far. The overarching project is to use causal models to define actual causation. In general, a definition of this kind – a *SEM definition of actual causation* – has two parts. The first is to give a recipe for how to read causal relations off a particular model. The second is to describe the domain of models over which the SEM definition quantifies. If a model belongs to this domain just in case it is *apt* for representing the target situation, then completing the second part of this project requires providing an account of *aptness*. This has been the initial task of this dissertation.

As a first point of clarification, I identified aptness as a two-part relation that holds between model-interpretation pairs and concrete situations. That is, it is the two-part relation that holds whenever that model-interpretation pair is of the right kind to represent the actual causal structure of that situation. I have therefore spoken of apt *model-interpretation pairs* throughout, rather than merely of apt *models*, while taking the target situation as given. I then developed the following account of this relation of aptness. First, define an *interpretation* of a model as an assignment of content to the variables of the model and a specification of modal profile. An interpretation of a model is *permissible* just in case it satisfies exclusivity, exhaustivity, and distinctness relative to the specified modal profile. A model-interpretation pair, $< \mathcal{M}_i, \mathcal{I}(\mathcal{M}_i) >$, will be *accurate* of its target situation just in case

$\mathcal{I}(\mathcal{M}_i)$ is *permissible*, the assignment, $\boldsymbol{\mathcal{A}}_{\mathcal{M}_i}$, is correct, and the entailed counterfactuals are true relative to the modal profile specified by $\mathcal{I}(\mathcal{M}_i)$. A model-interpretation pair, $< \mathcal{M}_i, \mathcal{I}(\mathcal{M}_i) >$, will be *apt* for representing its target situation just in case it satisfies Explicit Partial Mediation and is accurate of its target situation.

However, this account of aptness, when coupled with a definition of actual causation that existentially quantifies over all apt models, delivers counterintuitive results about what actually causes what. The Prince eating biscuits qualifies as an actual cause *simpliciter* of the plant dying, for example, in **The Prince and his Biscuits**. I have argued that the best response to this problem is the adoption of a *causal relativist* view of actual causation – a view whereby actual causation holds relative to a modal profile. It is a three-part relation, holding between a cause, an effect, and a background modal profile. On this view, the Prince eating biscuits is an actual cause of the plant dying only relative to the modal profile that holds fixed his character, the lock mechanism on the greenhouse, the biscuit holiday schedule, and the layout of the palace. Relative to the modal profile that permits these facts to vary, the Prince eating biscuits is *not* an actual cause of the plant dying.

I then went on to show how a causal relativist view of actual causation can defend strong proportionality against three objections from the literature. Finally, I concluded with an argument that, due to the difficulty of articulating aptness, an interventionist semantics for counterfactuals has not yet solved the importation problem.

While this discussion has placed causal relativism about actual causation on the table, much work remains to be done in articulating, situating, and motivating the view. Future research will require first precisifying the notion of modal profile and then explaining how a modal profile is determined. In particular, it remains to be seen whether and how *a situation* constrains the collection of modal profiles relative to which actual causal structures hold. The vignettes employed in Chapter 4 are carefully selected so that facts about the underlying situation justify consideration of a restricted modal profile. Sophie and Alice being in the scarlet-cyan factory yard justify consideration of a modal profile according to which the paint chips can only be scarlet or cyan. But their being in the factory yard doesn't seem to *demand* consideration of this restricted modal profile. There is a real sense in which – even in the factory yard – the chips *could have been* a non-scarlet-or-cyan color. Now consider, though, that Sophie and Alice are fluttering about the Home Depot paint section, which carries tens of thousands of paint samples. It's still certainly true that the chips could be many different non-scarlet-or-cyan colors in this new situation. But now there seems there would be no obvious justification for the more *restricted* modal profile, according to which the paint chips can only be scarlet or cyan. Does this lack of justification equate to a lack of a real possibility space? It's unclear. It seems natural to allow more *permissive* modal profiles in any given situation, but it is not as natural to allow more *restrictive* modal profiles – this only makes sense when some actual fact about the given situation entails the more restrictive modal profile. The question to be answered is whether this asymmetry in what strikes us as natural corresponds to an objective feature of the relationship between situations and their modal profiles.

It also remains to be seen in what ways a particular *causal inquiry* determines a modal profile, and whether natural language data about causal inquiries can support the claim that a well-formed inquiry implies a modal profile.

The answers to these questions will help to situate causal relativism about actual causation within theoretical space. In particular, they will serve to determine how causal relativism compares to contrastivism (see §4.6.c), and the ways in which it lines up with the context-sensitive view of actual causation defended by Menzies (2004a, 2004b, 2007).

Further motivation for a causal relativist view about actual causation is also called for. For example, a recent trend in the actual causation literature is to insist that a SEM definition needs to incorporate a distinction between default and deviant states of a system.[1] Arguably, the seeming need for such a distinction can be accommodated by Causal Relativism in the same way that it handles the "serious" qualification on exhaustivity. Intuitions commonly invoked in support of a default/deviant distinction can be explained as a preference for certain modal profiles over others. Additional support may be found if it could be shown that Causal Relativism adequately addresses infamous problems of causation involving transitivity and causation by omission. Unfortunately, I must leave the discussion here. As should be clear, there is more work to be done.

[1] See (Gallow, forthcoming; N. Hall, 2007; J. Halpern & Hitchcock, 2010; J. Y. Halpern, 2016b; J. Y. Halpern & Hitchcock, 2015; Menzies, 2017).

REFERENCES

Barker, S. (1999). Counterfactuals, Probabilistic Counterfactuals and Causation. *Mind*, *108*(431), 427–469.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Clarendon Press.

Blanchard, T. (2018). Explanatory Abstraction and the Goldilocks Problem: Interventionism Gets Things Just Right. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy030

Blanchard, T., & Schaffer, J. (2017). Cause Without Default. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation* (pp. 175–214). Oxford University Press. https://doi.org/10.1093/oso/9780198746911.003.0010

Bontly, T. D. (2005). Proportionality, causation, and exclusion. *Philosophia*, *32*(1–4), 331–348. https://doi.org/10.1007/BF02641629

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*(1), 139–166. https://doi.org/10.1007/s11098-012-9908-5

Cartwright, N. (2016). Single Case Causes: What is Evidence and Why. In *Philosophy of Science in Practice: Nancy Cartwright and the Nature of Scientific Reasoning* (pp. 11–24). Springer.

Ciardelli, I., Zhang, L., & Champollion, L. (2018). Two Switches in the Theory of Counterfactuals: A Study of Truth Conditionality and Minimal Change. *Linguistics and Philosophy*, *41*, 577–621.

Clarke-Doane, J. (2020). *Morality and Mathematics*. Oxford University Press.

Donne, J. (1624). *Devotions upon Emergent Occasions*.

Dretske, F. (1977). Referring to Events. *Midwest Studies in Philosophy*, *2*(1), 90–99.

Edgington, D. (2004). Counterfactuals and the Benefit of Hindsight. In P. Dowe & P. Noordhof (Eds.), *Cause and Chance: Causation in an Indeterministic World* (pp. 12–27). Routledge.

Elga, A. (2001). Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science*, *68*(3), 313–324.

Fine, K. (1975). Critical Notice: Counterfactuals. *Mind*, *84*, 451–458.

Franklin-Hall, L. R. (2016). High-Level Explanation and the Interventionist's 'Variables Problem.' *The British Journal for the Philosophy of Science*, *67*(2), 553–577. https://doi.org/10.1093/bjps/axu040

Galles, D., & Pearl, J. (1998). An Axiomatic Characterization of Causal Counterfactuals. *Foundations of Science*, *3*(1), 151–182.

Gallow, J. D. (2016). A Theory of Structural Determination. *Philosophical Studies*, *173*(1), 159–186.

Gallow, J. D. (forthcoming). A Model-Invariant Theory of Causation. *Philosophical Review*.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Harvard University Press.

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press. http://www.hup.harvard.edu/catalog.php?isbn=9780674852716

Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, *132*(1), 109–136. https://doi.org/10.1007/s11098-006-9057-9

Hall, Ned. (2006). Structural Equations and Causation. *Manuscript*.

Halpern, J. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, *12*, 317–337.

Halpern, J., & Hitchcock, C. (2010). Actual Causation and the Art of Modeling. In *Causality, Probability, and Heuristics: A Tribute to Judea Pearl* (pp. 383–406). London: College Publications.

Halpern, J. Y. (2016a). *Actual Causality*. MIT Press. http://www.jstor.org.ezproxy.gc.cuny.edu/stable/j.ctt1f5g5p9

Halpern, J. Y. (2016b). Appropriate Causal Models and the Stability of Causation. *The Review of Symbolic Logic*, *9*(01), 76–102. https://doi.org/10.1017/S1755020315000246

Halpern, J. Y., & Hitchcock, C. (2015). Graded Causation and Defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413–457.

Halpern, & Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887. https://doi.org/10.1093/bjps/axi147

Hausman, D. M., Stern, R., & Weinberger, N. (2014). Systems without a graphical causal representation. *Synthese*, *191*(8), 1925–1930.

Hawthorne, J. (2005). Chance and Counterfactuals. *Philosophy and Phenomenological Research*, *70*(2), 396–405.

Hiddleston, E. (2005a). A Causal Theory of Counterfactuals. *Nous*, *39*(4), 232–257.

Hiddleston, E. (2005b). Causal Powers. *British Journal for the Philosophy of Science*, *56*, 27–59.

Hitchcock, C. (1996a). Farewell to Binary Causation. *Canadian Journal of Philosophy*, *26*, 267–282.

Hitchcock, C. (1996b). The Role of Contrast in Causal and Explanatory Claims. *Synthese*, *107*(3), 395–419.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy*, *98*(6), 273–299. https://doi.org/10.2307/2678432

Hitchcock, C. (2004). Routes, Processes, and Chance-Lowering Causes. In P. Dowe & P. Noordhof (Eds.), *Cause and Chance: Causation in an Indeterministic World*. Routledge.

Hitchcock, C. (2007a). Prevention, Preemption, and the Principle of Sufficient Reason. *The Philosophical Review*, *116*(4), 495–532.

Hitchcock, C. (2007b). What's Wrong with Neuron Diagrams? In J. K. Campbell, M. O'Rourke, & H. S. Silverstein (Eds.), *Causation and Explanation* (pp. 4–69). MIT Press.

Hitchcock, C. (2009). Causal Modelling. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford Handbook of Causation* (Vol. 1, pp. 299–314). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199279739.003.0015

Hitchcock, C. (2011). Trumping and Contrastive Causation. *Synthese*, *181*(2), 227–249.

Hitchcock, C. (2012). Events and Times: A Case Study in Means-Ends Metaphysics. *Philosophical Studies*, *160*(1), 79–96.

Hitchcock, C. (2018). Causal Models. *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2018/entries/causal-models/>.

Hoffmann-Kolss, V. (2014). Interventionism and Higher-level Causation. *International Studies in the Philosophy of Science*, *28*(1), 49–64. https://doi.org/10.1080/02698595.2014.915653

Huber, F. (2013). Structural Equations and Beyond. *Review of Symbolic Logic*, *6*(4), 709–732.

Kaufmann, S. (2013). Causal Premise Semantics. *Cognitive Science*, *37*(6), 1136–1170.

Kim, J. (1974). Noncausal Connections. *Noûs*, *8*(1), 41–52. https://doi.org/10.2307/2214644

Kvart, I. (1986). *A Theory of Counterfactuals*. Hackett Publishing.

Lewis, D. (1973a). *Counterfactuals*. Harvard University Press.

Lewis, D. (1973b). Counterfactuals and Comparative Possibility. *Journal of Philosophical Logic*, *2*(4), 418–446.

Lewis, D. (1973c). Causation. *The Journal of Philosophy*, *70*(17), 556. https://doi.org/10.2307/2025310

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Nous*, *13*(4), 455–476.

Lewis, D. (1986). Postscripts to "Counterfactual Dependence and Time's Arrow." In *Philosophical Papers: Volume II* (pp. 52–66). Oxford University Press.

Lewis, D. (2000). Causation as Influence. *The Journal of Philosophy*, *97*(4), 182. https://doi.org/10.2307/2678389

List, C., & Menzies, P. (2009). Nonreductive Physicalism and the Limits of the Exclusion Principle: *Journal of Philosophy*, *106*(9), 475–502. https://doi.org/10.5840/jphil2009106936

Maslen, C. (2004). Causes, Contrasts, and the Nontransitivity of Causation. In Ned Hall, L. A. Paul, & J. Collins (Eds.), *Causation and Counterfactuals* (pp. 341–357). Cambridge: Mass.: Mit Press.

Maslen, C. (2017). Pragmatic Explanations of the Proportionality Constraint on Causation. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation* (pp. 58–72). Oxford University Press.

McDonnell, N. (2017). Causal exclusion and the limits of proportionality. *Philosophical Studies*, *174*(6), 1459–1474. https://doi.org/10.1007/s11098-016-0767-3

McDonnell, N. (2018). Transitivity and proportionality in causation. *Synthese*, *195*(3), 1211–1229. https://doi.org/10.1007/s11229-016-1263-1

Menzies, P. (2004a). Difference Making in Context. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and Counterfactuals* (pp. 341–367). Oxford University Press.

Menzies, P. (2004b). Causal Models, Token Causation, and Processes. *Philosophy of Science*, *71*(5), 820–832. https://doi.org/10.1086/425057

Menzies, P. (2007). Causation in Context. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.

Menzies, P. (2008). The Exclusion Problem, The Determination Relation, and Contrastive Causation. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced—New Essays on Reduction, Explanation, and Causation* (pp. 196–217). Oxford University Press.

Menzies, P. (2017). The Problem of Counterfactual Isomorphs. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press.

Menzies, P., & List, C. (2010). The Causal Autonomy of the Special Sciences. In C. Mcdonald & G. Mcdonald (Eds.), *Emergence in Mind*. Oxford University Press.

Montminy, M., & Russo, A. (2016). A Defense of Causal Invariantism. *Analytic Philosophy*, *57*(1), 49–75.

Northcott, R. (2008). Causation and Contrast Classes. *Philosophical Studies*, *139*, 111–123.

Papineau, D. (2013). Causation is Macroscopic but Not Irreducible. In S. C. Gibb & R. Ingthorsson (Eds.), *Mental Causation and Ontology* (p. 126). Oxford University Press.

Paul, L. A. (2000). Aspect Causation. *The Journal of Philosophy*, *97*(4), 235. https://doi.org/10.2307/2678392

Paul, L. A., & Hall, N. (2013). *Causation: A User's Guide*. Oxford University Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference* (Second edition., 3rd printing..). Cambridge University Press.

Pearl, J. (2013). Structural Counterfactuals: A Brief Introduction. *Cognitive Science*, *37*, 977–985.

Priest, G. (2018). Some New Thoughts on Conditionals. *Topoi*, *37*(3), 369–377. https://doi.org/10.1007/s11245-016-9438-4

Reiss, J. (2013a). Contextualising Causation Part 1. *Philosophy Compass*, *8*(11), 1066–1075.

Reiss, J. (2013b). Contextualising Causation Part II. *Philosophy Compass*, *8*(11), 1076–1090.

Santorio, P. (2014). Filtering Semantics for Counterfactuals: Bridging Causal Models and Premise Semantics. In T. Snider, S. D'Antonio, & M. Wiegand (Eds.), *Semantics and Linguistic Theory* (pp. 494–513). LSA and CLC Publications.

Santorio, P. (2019). Interventions in Premise Semantics. *Philosophers' Imprint*, *19*(1).

Sartorio, C. (2010). The Prince of Wales problem for counterfactual theories of causation. In A. Hazzlett (Ed.), *New Waves in Metaphysics* (pp. 259–276). Palgrave Macmillan.

Schaffer, J. (2004). Counterfactuals, Causal Independence, and Conceptual Circularity. *Analysis*, *64*(4), 299–309.

Schaffer, J. (2005). Contrastive Causation. *Philosophical Review*, *114*(3), 327–358. https://doi.org/10.1215/00318108-114-3-327

Schaffer, J. (2010). Contrastive Causation in the Law. In *Legal Theory* (Vol. 16, pp. 259–297). Cambridge University Press.

Schaffer, J. (2012). Causal Contextualism. In M. Blaauw (Ed.), *Contrastivism in Philosophy: New Perspectives*. Routledge.

Schaffer, J. (2016). The Metaphysics of Causation. *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/>

Schroeter, L. (2019). Two-Dimensional Semantics. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2019/entries/two-dimensional-semantics/>

Schulz, K. (2011). If You'd Wiggled A, Then B Would've Changed. *Synthese*, *179*, 239–251.

Shapiro, L., & Sober, E. (2012). Against proportionality. *Analysis*, *72*(1), 89–93. https://doi.org/10.1093/analys/anr135

Sinnott-Armstrong, W. (2021). Contrastive Mental Causation. *Synthese*, *198*(S.I.: Materialism and Metaphysics), 861–883.

Slote, M. (1978). Time in Counterfactuals. *Philosophical Review*, *87*, 3–27.

Spirtes, P., Glymour, C. N., & Scheines, R. (1993). *Causation, prediction, and search*. Springer-Verlag.

Stalnaker, R. (1968). A Theory of Conditionals. *American Philosophical Quarterly*, 98–112.

Starr, W. (2019). Counterfactuals. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2019/entries/counterfactuals/>

Steglich-Petersen, A. (2012). Against the Contrastive Account of Singular Causation. *The British Journal for the Philosophy of Science*, *63*(1), 115–143. https://doi.org/10.1093/bjps/axr024

Wasserman, R. (2006). The Future Similarity Objection Revisited. *Synthese*, *150*(1), 57–67.

Weslake, B. (2013). Proportionality, Contrast and Explanation. *Australasian Journal of Philosophy*, *91*(4), 785–797. https://doi.org/10.1080/00048402.2013.788045

Weslake, B. (2015). A Partial Theory of Actual Causation. *British Journal for the Philosophy of Science*.

Weslake, B. (2017). Difference-making, Closure, and Exclusion. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation* (Vol. 1, pp. 215–231). Oxford University Press. https://doi.org/10.1093/oso/9780198746911.003.0011

Weslake, B. (forthcoming). Exclusion Excluded. *International Studies in the Philosophy of Science*.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, J. (2006). Sensitive and Insensitive Causation. *Philosophical Review*, *115*(1), 1–50.

Woodward, J. (2008). Mental Causation and Neural Mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation* (pp. 218–262). Oxford University Press.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318. https://doi.org/10.1007/s10539-010-9200-z

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, *91*(2), 303–347. https://doi.org/10.1111/phpr.12095

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072. https://doi.org/10.1007/s11229-015-0810-5

Woodward, J. (2018). Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance. *Synthese*, 1–29.

Yablo, S. (1992). Mental Causation. *The Philosophical Review*, *101*(2), 245–280. https://doi.org/10.2307/2185535

Yablo, S. (2003). Causal Relevance. *Philosophical Issues*, *13*(1), 316–328.

Yang, E. (2013). Eliminativism, interventionism and the Overdetermination Argument. *Philosophical Studies*, *164*(2), 321–340. https://doi.org/10.1007/s11098-012-9856-0