*Forthcoming in Philosophy Compass*

# Causal Models and Metaphysics – Part 2: Interpreting Causal Models[1]

Jenn McDonald

> [C]ausal models are mathematical representations of concrete situations, and whenever one indulges in representation the question may arise as to whether the representation is faithful to reality. (Blanchard & Schaffer, 2017, p. 181)

> [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)

**Abstract:** This paper addresses the question of what constitutes an apt interpreted model for the purpose of analyzing causation. I first collect universally adopted aptness principles into a basic account, flagging open questions and choice points along the way. I then explore various additional aptness principles that have been proposed in the literature but have not been widely adopted, the motivations behind their proposals, and the concerns with each that stand in the way of universal adoption. I conclude that the remaining work of articulating aptness for a SEM analysis of causation is tied up with issue to do with modality, ontology, and mereology. Continuing this work is therefore likely to shed light on the relationship between these areas and causation more generally.

## §1 Introduction

A causal model is not apt on its own. As a formal object, a model has no real-world content unless interpreted. An interpreted causal model is apt, but only in relation to some situation(s) under inquiry. So, a model is apt only *under an interpretation* and only *relative*

[1] This paper was greatly improved by discussion with and feedback from (in alphabetical order): Justin Clarke-Doane, Christopher Hitchcock, David Papineau, James Woodward, and Tomasz Wysocki.

*to a target situation or set of situations*. Thus, aptness is a relation between three things – a model, an interpretation, and a situation (or set of such). In fact, it will also be relative to an aim or purpose. Aptness as it's discussed in this paper is designed for the provision of a metaphysical analysis of causation.

## §2 Accuracy

Everyone has an aptness problem. It won't do to quantify over just any interpreted models, since at minimum they will need to be *accurate* – that is, an interpreted model will need to get the target right. I take an "accurate" interpreted model to be one that says only true things about its target. There is room, perhaps, to argue that this is overly demanding, possibly counterproductive. Representations that idealize, approximate, or in any event get close enough to the truth are often as useful as – even, the argument might go, more useful than – accurate ones (Elgin, 2004; Potochnik, 2017). However, I will set this line of argument aside. This paper will assume, in accord with the universal position taken in the SEM literature relevant to the metaphysics of causation, that an apt interpreted model is, at minimum, accurate. So, what does an interpreted model say, and when is that true?

On the method of interpretation from Part 1, an interpreted nonspecific SEM says two things: its interpretation satisfies the permissibility conditions, and the relations represented by its equations really hold. It is accurate insofar as this is true. More exactly:

**Accuracy – GC$_S$** A causal model, $\mathcal{M}_i$, is accurate of a set of situations, $\mathbb{S}$, on an interpretation $\mathcal{I}(\mathcal{M}_i)$, just in case …

i. $\mathcal{I}(\mathcal{M}_i)$ is a permissible interpretation of $\mathcal{M}_i$ for representing every $\mathbb{s} \in \mathbb{S}$; and

ii. The relations represented by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$ on $\mathcal{I}(\mathcal{M}_i)$ hold in every $\mathbb{s} \in \mathbb{S}$.[2]

[2] As discussed in Part 1, $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$ is the linkage of a model, $\mathcal{M}_i$.

A specific SEM under either a general or particular interpretation says, in addition, that the property instances represented by the values assigned to the exogenous variables actually occur. More exactly:

**Accuracy – AC$_S$** A causal model, $\mathcal{M}_i$, is accurate of a given situation, $\mathbb{s}$, on an interpretation, $\mathcal{I}(\mathcal{M}_i)$, just in case …

i. $\mathcal{I}(\mathcal{M}_i)$ is a permissible interpretation of $\mathcal{M}_i$ for representing $\mathbb{s}$;

ii. The relations represented by $\boldsymbol{\mathcal{L}}_{\mathcal{M}_i}$ on $\mathcal{I}(\mathcal{M}_i)$ hold in $\mathbb{s}$; and

iii. The property instances represented by $\boldsymbol{U}_{\mathcal{M}_i}$ given $\boldsymbol{\mathcal{A}}_{\mathcal{M}_i}$, on $\mathcal{I}(\mathcal{M}_i)$ occur in $\mathbb{s}$.[3]

Note that these are merely schemata, for two reasons. First and to be taken up shortly, what constitutes permissibility of an interpretation is yet to be specified. Second and to be taken up in §4, no particular view about what the equations represent is assumed. Accuracy simply requires that whatever relations represented by an equation really do hold. Yet, as we'll see, the details vary depending on what these are.

## §3 Permissible Interpretations

As discussed in Part 1, it is universally agreed that an interpretation must satisfy exclusivity, exhaustivity, and distinctness. That is, any two property instances mapped to any two values of the same variable are mutually exclusive in the target situation or set of situations ("exclusivity"), the range of property instances mapped to the full set of values of a given variable are jointly exhaustive ("exhaustivity"), and any two property instances mapped to values of different variables are "distinct," or independent of each other ("distinctness"). But whether these are satisfied depends on what it is for two (or more) property instances to be exclusive, exhaustive, or distinct.

### §3.1 Exclusivity and Distinctness

[3] As discussed in Part 1, $\boldsymbol{U}_{\mathcal{M}_i}$ is the set of exogenous variables of $\mathcal{M}_i$, and $\boldsymbol{\mathcal{A}}_{\mathcal{M}_i}$ is $\mathcal{M}_i$'s assignment of values to these variables.

Take exclusivity and distinctness first. Since either can be satisfied independently of the other, these conditions aren't exactly corollaries. But they track the same thing – whether two property instances can possibly co-occur. If they can, then they are distinct and not exclusive. If they cannot, they are exclusive and not distinct. Whether two property instances could co-occur is a function of the relationship between the respective objects, properties, and time periods, as well as what counts as possible. So, what counts? This question is primarily couched in terms of kinds of modalities: logical, metaphysical, conceptual, etc. Blanchard and Schaffer, for example, characterize possibility in terms of "logical [and] metaphysical relations." (2017, p. 182) Woodward more fully characterizes possibility "in terms of [the] assumed definitional, logical, mathematical, mereological or supervenience relations." (2015, p. 316) Some invoke the sophisticated account of distinctness given by Lewis (1986).[4] While sophisticated, however, the account is incomplete.

It thus remains an open question how these details should be filled in. However they are, the correct characterization of possibility arguably goes beyond fixing on a kind of modality. It seems that what counts as possible additionally depends on something further. If there's only a single train travelling down the tracks, then whether the left-hand track is occupied fails to be distinct from whether the right-hand track is occupied (during the same time period). The one train cannot travel on both tracks at once. But there could have been another train. Had there been, then both tracks could simultaneously be occupied. Given this, each track being occupied is distinct from the other. The accurate representation of a given situation or set of situations, then, depends in part on what features are permitted to vary. This, in turn, determines what counts as an actual cause since a SEM analysis quantifies over apt interpreted models (see Part 1). If the number of trains is fixed, then the tracks being occupied are exclusive, and my having switched the lever to direct the train down the right-hand track is not an actual cause of its arriving at the station. If the number of trains can vary, then the tracks being occupied are distinct, and my switching the lever is an actual cause. So, how many trains could there be?

[4] See, for example, (Hitchcock, 2004, p. 146, 2007, p. 502).

The latter representation is allowable on the grounds that it is logically (metaphysically/ conceptually/ etc.) possible for there to be more than one train. However, allowing this contravenes common causal intuition, which says that my flipping the lever is not a cause. After all, the train arrives at the station either way![5] There is a tension, then, between what's been said about exclusivity and distinctness, and what intuition deems a cause.

The clearest way to think about this, to my mind, is to relativize the satisfaction of exclusivity and distinctness to a further parameter: a setting of background possibilities. Though I won't argue it here, there is reason to think this relativity holds of exhaustivity and of whether the relations represented by the equations really hold, as well.[6] Since this profiles a situation in terms of its modal character – of how a situation might have gone – call it a "modal profile." A simple way to understand the modal profile of a *situation* is on analogy with the modal profile of an event or of an object: it answers the question of how this could have varied while still remaining the same, in some important sense. Different modal profiles answer the question differently. Is there a correct answer? Are there at least wrong answers?

Perhaps what's possible can in some way be given by what holds in the (fully-specified) target situation. However this might go, it can't be the whole story. This is because modal profiles are restricted in part by representational choices. Certain representations come with presuppositions, restricting what the interpreted model can treat as possible. Suppose a binary variable, $X$, represents a child being under or equal to the height of 36 inches or over 36 inches. This range of properties presupposes that the child exists. The interpreted model is therefore incapable of representing the possibility of the child not existing, given distinctness. Such a restriction will be reflected in the modal profile.

The introduction of modal profiles renders the above accuracy schemata incomplete. An interpretation is now permissible, and relations hold of a situation, only relative to a modal

[5] For discussion of this switching example, see (J. Y. Halpern, 2016a, p. 38; Woodward, 2016, p. 1063).

[6] But see (McDonald, 2022).

profile. So, an interpreted model is accurate of its target only relative to a modal profile. Relative to one that allows for the possibility of there being more than one train, two variables representing the separate tracks being occupied satisfies distinctness. But relative to a modal profile that restricts the number of trains to the one actual train, distinctness is violated. This relativity is plausibly what Woodward has in mind above when he references the "assumed" relations of dependence.[7]

Of course, one could reject the introduction of modal profiles as the best explication of the variability gestured at above. Alternatively, a precise account of the nature of causal relata might entail independently justified answers to the foregoing questions: What counts as exclusive or distinct? How could the target situation possibly have gone? Either way, this issue connects SEM analyses of causation back up with traditional debates in metaphysics about mereology, ontology, essentialism, events, objects, etc.[8] As it turns out, then, causal models are less neutral on the question of causal relata than they first appear.

### §3.2 Independent Manipulability

A related candidate condition on permissibility is that of "Independent Manipulability," or "Independent Fixability," which requires that any combination of values be possible (Weslake, forthcoming; Woodward, 2008, 2015, p. 316, 2016, p. 1054; Yang, 2013; Zhong, 2020). That is, any variable taking any one of its values be compossible with every setting of values to all the other variables. Prima facie, this follows from distinctness insofar as the matter of what counts as possible is settled in the same way for both conditions.

### §3.3 Exhaustivity and Serious Possibilities

[7] Woodward is generally sensitive to a kind of relativity to background conditions (2003), although he carefully refrains from making metaphysical pronouncements. Similar sensitivity can be found in (Gallow, 2016; Menzies, 2004b; Statham, 2018), as well as in the discussion around causation as a contrastive relation (Hitchcock, 1996b, 1996a, 2011; Maslen, 2004; Northcott, 2008; Schaffer, 2005, 2012; Steglich-Petersen, 2012).

[8] See, for example, (Casati & Varzi, 2000, 2023).

The third universally accepted condition is exhaustivity, which demands that no possible alternative property instance be left out. More exactly, a range of property instances, $\{p_1(o)_t,\ p_2(o)_t, \ldots, p_n(o)_t\}$, will count as jointly exhaustive in a situation (or set of situations) when there is no property, $p_i$, possibly instantiated by the underlying object, $o$, (or set of objects, $\boldsymbol{O}$,) during the given time period, $t$, that would exclude $o$ (or $o \in \boldsymbol{O}$) instantiating any of the properties $\{p_1, p_2, \ldots, p_n\}$ at $t$. This raises the same question as to what counts as possible. Suppose a variable represents the color of a bell pepper. Need it represent all colors on the visible spectrum in order to be exhaustive? Or, could it count as exhaustive by only representing the colors possibly manifest in a bell pepper – that is, purple, green, yellow, orange, red? Then, again, is the visible spectrum broad enough?

A related, but controversial, condition is "serious possibilities," which demands that an interpreted model represent only the serious, genuine, or relevant possibilities (Blanchard & Schaffer, 2017, p. 182; Fenton-Glynn, 2021, p. 46; Hitchcock, 2001, p. 287; Woodward, 2016, p. 1064; Wysocki, 2023, p. 3537).[9] Then, a range of property instances is exhaustive only if it covers all serious (/genuine/relevant) alternatives.

A further application of this condition is how it handles the problem of causation by omission. On a traditional counterfactual account, the Queen's failing to water the plants causes them to die, despite her having nothing to do with them. This isn't quite right. But if her watering the plants is not a serious (/genuine/relevant) possibility, then it cannot be aptly represented by an interpreted model. Problem solved.

What counts as serious, genuine, or relevant? Crucially, the nature of the guiding principles here will determine whether this condition is constitutively determined by us. If it is, then a realist construal of actual causation is off the table. A realist about actual causation who endorses this condition must deliver a purely mind and language independent set of guiding principles. But one could be an antirealist about *actual causation*, and yet a realist about some other (perhaps more basic) notion of causation, such as the underlying *token causal*

[9] Lewis (2000) raises a similar condition in a pure counterfactual context.

*structure* (Hitchcock, 2003, 2007; Kuorikoski, 2014). Given such a view, serious possibilities – as a condition on actual causation – could be determined pragmatically, say, without undermining the mind and language independence of causation more generally.

## §4 Equations and Interventions

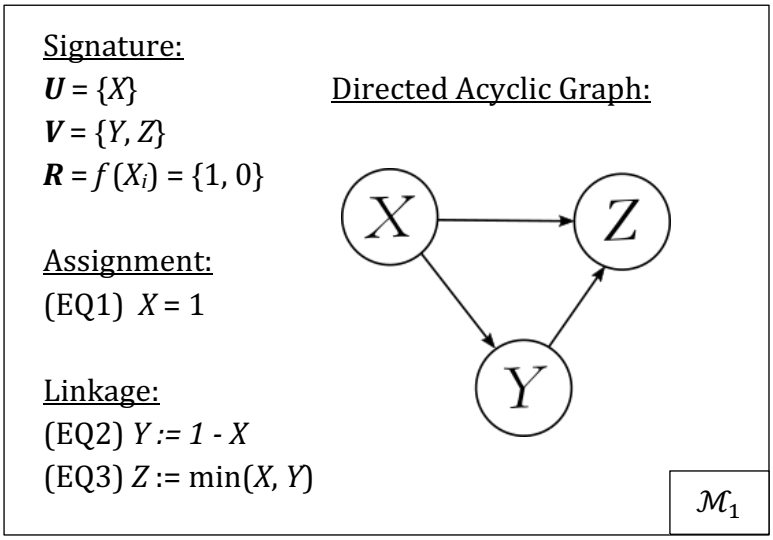
What else is required for an interpreted model to be apt depends on one's view about what the equations represent. This includes what an intervention represents, if anything. This section surveys various options. One thing to keep in mind is that the following views about causal metaphysics can pair up with views about causal epistemology in various ways. For example, one might take *F*-dependencies as more metaphysically basic (than *G*-dependencies) but *G*-dependencies as more epistemically basic (than *F*-dependencies). Thus, while equations represent *F*-dependencies when the project is metaphysical, they represent *G*-dependencies when the project is epistemological. Again, I focus on the metaphysics.

### §4.1 Causal Primitives and Regularities

Some views treat equations as representing causal influence taken as primitive (Cartwright, 2016; Gallow, 2016, 2021, 2023), reducing actual causation to causal influence (see Part 1, §5). But some views take this further, reducing causal influence to lawlike regularities or nomological dependencies (Andreas & Gunther, forthcominga, forthcomingb; Papineau, 2022; Pearl, 2000/2009).[10] In any case, an advantage here is the reduction of (at least some kinds of) counterfactual dependence to causal influence (Briggs, 2012; Galles & Pearl, 1998; Hiddleston, 2005; Pearl, 2000/2009; Starr, 2019). How such a semantics of counterfactuals measures up against others is an ongoing research question.

[10] Whether Pearl belongs here or in the counterfactual camp depends on how he sees the relationship between the nature of a law and its empirical content. He takes the functions of each equation to represent scientific laws whose empirical content is a set of "generic, counterfactual relationships among [ordered sets of property instances] that are applicable to every hypothetical scenario." (2000/2009, p. 310) See also (Pearl, 2000/2009, p. 160: fn17; 165; §7.2.2).

Primitive and regularity views treat equations as representing type-level relations regardless of whether a nonspecific or specific SEM is at hand, and of whether a general or particular interpretation is utilized. The truth conditions for the equations will essentially be the same in any case. As an example, consider the following particular interpretation of a specific SEM.




$\mathcal{I}(\mathcal{M}_1)_P$: Population = $\{< Suzy,\ Billy,\ window >\}$

$$X\ (Suzy) = \begin{cases} 1\ if\ throws\ a\ rock \\ 0\ if\ doesn't\ throw\ a\ rock \end{cases}$$

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

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

On primitive views, the equation '$Z$ := max($X$, $Y$)' under $\mathcal{I}(\mathcal{M}_1)_P$ represents the relation of causal influence linking rocks being thrown (as opposed to rocks not being thrown) to window shatterings (as opposed to windows remaining intact). Alternatively, the equation

might represent a regularity or nomological dependency that holds between rock throwings and window shatterings. The interpreted model is accurate only if these relations really do hold of the target situation.

### §4.2 Counterfactual Dependencies

Most views treat equations as representing counterfactual dependencies of some kind. How the details shake out depends further on whether the interpretation is general or particular. The philosophy literature is primarily concerned with particular interpretations. This is likely due to a focus on deterministic actual causation coupled with an implicit assumption that type-level causation supervenes on actual. So, take the particular first. Consider $< \mathcal{M}_1, \mathcal{I}(\mathcal{M}_1)_P >$, from before. The prominent view in this camp treats the system of equations {EQ1, EQ2, EQ3} as representing complex counterfactuals about particulars (Hall, 2007; J. Y. Halpern, 2016a; Handfield et al., 2008; Hitchcock, 2001, 2007; Kroedel, 2019). On this view, the equation '$Z := \max(X, Y)$' under $\mathcal{I}(\mathcal{M}_1)_P$ would represent the following counterfactuals: 'had Suzy thrown a rock, then the window would have shattered'; 'had Billy thrown a rock, then the window would have shattered'; and 'had neither Suzy nor Billy thrown a rock, then the window would have not shattered.' The interpreted model is accurate, then, only if these counterfactual dependencies are true of the target situation.

Of course, this requires specifying a semantics with which to evaluate the counterfactuals, and there is disagreement over which one. If the semantics can be given without invoking causal relations, then this view is amenable to reducing actual causation to counterfactual dependence (Hall, 2007; J. Y. Halpern, 2016a; Handfield et al., 2008; Hitchcock, 2001, 2007; Kroedel, 2019; Pearl, 2000/2009, p. 310).[11] If it can't, it is still amenable to giving an illuminating albeit non-reductive account of actual causation in terms of both counterfactuals and other relations of causal relevance (Woodward, 2003). Where

[11] It should be noted that Handfield et al. (2008) require a further aptness condition that every parenthood relation also correspond to a physical process. They therefore reduce actual causation to counterfactual dependencies that are specifically underwritten by physical processes.

mentioned, it seems many further treat general causation as reducible to actual causation (Woodward, 2003, p. 40), although a SEM account of this has yet to be attempted.

General interpretations are slightly different. Consider:

$\mathcal{I}(\mathcal{M}_1)_G$: Population = $\{< Suzy,\ Billy,\ window >, < Jamie,\ Bobby,\ bottle >\}$

$$X\left(\begin{smallmatrix} Suzy \\ Jamie \end{smallmatrix}\right) = \begin{cases} 1\ if\ throws\ a\ rock \\ 0\ if\ doesn't\ throw\ a\ rock \end{cases}$$

$$Y\left(\begin{smallmatrix} Billy \\ Bobby \end{smallmatrix}\right) = \begin{cases} 1\ if\ throws\ a\ rock \\ 0\ if\ doesn't\ throw\ a\ rock \end{cases}$$

$$Z\left(\begin{smallmatrix} window \\ bottle \end{smallmatrix}\right) = \begin{cases} 1\ if\ shatters \\ 0\ if\ doesn't\ shatter \end{cases}$$

'$Z$ := max($X$, $Y$)' under $\mathcal{I}(\mathcal{M}_1)_G$ would thus represent: 'had Suzy thrown a rock, then this window would have shattered'; 'had Jamie thrown a rock, then that bottle would have shattered'; 'had Billy thrown a rock, then this window would have shattered'; 'had Bobby thrown a rock, then that bottle would have shattered'; 'had neither Suzy nor Billy thrown a rock, then this window would not have shattered'; and 'had neither Jamie nor Bobby thrown a rock, then that bottle would not have shattered.' Only if these are true of the target situation(s) will the interpreted model be accurate.

### §4.3 On the Notion of Intervention

Part 1 defines an intervention formally as an operation on a model: an "intervention" on a model, $\mathcal{M}_{X=x}$, replaces the $X$-equation in $\mathcal{M}$ with the constant equation $X = x$. But what does this operation represent? In fact, for causal primitive views, an intervention need not represent anything in the underlying metaphysics. To see this, note that the essential role of an intervention is to ensure asymmetry and circumvent the confounding that results from correlations, regularities, and counterfactual dependencies holding between (distinct) non-

causes. For example, a regularity will hold between two joint effects of a common cause, and one may counterfactually depend on the other. Yet, neither is a cause of the other. For primitive views, confounding is straightforwardly identified.

But the need to distinguish between spurious and genuine causal relations is a well-known problem incurred by regularity and counterfactual analyses. For a SEM regularity analysis, one might handle this by somehow extending a proposal from Papineau (2022). He argues that, given probability distributions over possible assignments to the exogenous variables (an additional component in considering probabilistic SEMs), requiring that exogenous variables be probabilistically independent will entail an order on a system of equations. Alternatively, a directionality could be built directly in to the regularities (Andreas & Gunther, forthcominga, forthcomingb). For neither of these options does an intervention need to represent something in the world. But tying an intervention to something in the world may be a third way to avoid confounding.

What about a counterfactual view? There are myriad ways to alter a situation so that it now involves the property-instance given by the antecedent in place of the original one. But not all such alterations are relevant to causation. Consider the reading of a barometer. The reading can be altered by adjusting the air pressure or by interfering with the barometer's mechanism. Adjusting the air pressure will affect the occurrence of the storm. But this fails to show that the barometer causes the storm. We've simply altered a common cause. Consider the counterfactual, 'had the barometer read a higher pressure, then there would not have been a storm.' This counterfactual is satisfied by a world in which the higher reading of the barometer is brought about by higher pressure. Insofar as this world is relevant to its evaluation, the counterfactual is true. But the truth of this counterfactual doesn't correspond to a direct causal relation. These "backtracking counterfactuals" – ones whose truth requires that changes occur before the time of the antecedent – are a problem for identifying causal relations. The challenge is how to rule them out not simply by fiat.[12]

[12] To do so by fiat is both ad hoc and would build temporal order into causal order – eliminating the possibility of reducing the former to the latter and ipso facto ruling out backwards causation.

This is where an intervention comes in. The question of what an intervention represents, in this context, is a question about how a counterfactual semantics should restrict the possible ways in which the antecedent may have been brought about. This matter is contentious. Hall (2007) argues that failure to attend to this issue has led to inapt models and muddied the discussion. Woodward (2003) argues that his worldly notion of an intervention is superior to that of the traditional notion of 'miracles', in part due to the issue of 'early' versus 'late' miracles.[13] Glynn (2013) counters this, arguing that miracles work just fine. This strikes me as a key question whose import outstrips the amount of attention it has received. Despite the shift in terminology from a traditional counterfactual analysis to current SEM analyses, it poses the same philosophical challenge. Among other things, the stakes involve whether a reductive analysis of causation in terms of counterfactuals is possible.[14]

## §5 Additional Principles of Variable Selection

Beyond the accuracy schemata and detail so far filled in, no principle of variable selection yet enjoys universal acceptance. This section discusses additional proposals, including further conditions on permissible interpretations and conditions governing the choice of what to represent. These latter conditions govern, for example, the introduction of new or elimination of existing variables, enabling or removing certain representational choices.

### §5.1 Intrinsic Characterizations and Naturalness

One condition on permissible interpretations that has been proposed requires that whatever values represent be intrinsically characterized. So, values must only represent property

[13] For what I'm calling the traditional view, see (Lewis, 1973b, 1973c, 1973a, 1979).

[14] Indeed, there is independent reason to think a reductive account isn't viable. Many problem cases suggest a similarity semantics based purely on miracles cannot work, and that causation must be invoked in order to make sense of counterfactual intuitions (Edgington, 2004; Elga, 2001; Fine, 1975; Schaffer, 2004; Wasserman, 2006). Note that such cases mostly, but not exclusively, involve indeterminism.

instances whose constitutive properties are intrinsic (Blanchard & Schaffer, 2017, p. 182; Fenton-Glynn, 2021, p. 45; Menzies, 2004b, 2004a). Otherwise, counterintuitive verdicts can be generated. Consider the fact of Socrates's death. Suppose we model it with two variables – one that represents my instantiating the property of being such that Socrates died in Athens in 399 B.C. or not so instantiating, and one that represents Plato instantiating the property of grieving his teacher or not. Given such an interpreted model, my being such that Socrates died would count as an actual cause of Plato grieving, satisfying any extant SEM recipe. Backwards causation aside, this is strange. It clearly fails to capture what causation is. Happily, an interpreted model like this would be inapt given an intrinsicality requirement.

However, 'intrinsic' may be too strong. It would also rule out a variable that represents Xanthippe instantiating the property of becoming a widow or not so instantiating. But surely this property can have causal force. Suppose Xanthippe held little love for Socrates, but values being a married woman. Then, it is not Socrates's death per se that makes her sad, but her becoming a widow. Yet, an intrinsicality requirement would forbid such representation.

Better, perhaps, would be a restriction against purely extrinsic properties. This assumes that properties can be reasonably positioned on a spectrum from purely intrinsic to purely extrinsic. Being a widow is plausibly more intrinsic than being such that Socrates died in Athens in 399 B.C. So, it survives the purge, while properties like being such that Socrates died in Athens in 399 B.C. are ruled out by a prohibition on purely extrinsic properties.

Alternatively, one might require that values represent only natural properties, on some criterion of naturalness. The criterion cannot be overly strict for reasons similar to what was just argued – otherwise, the overall analysis will grossly under-generate causes. A looser restriction like 'no gerrymandered or grue-some properties' would likely fit better with causal judgment, but at the cost of a vaguer criterion.

### §5.2 Proportionality

Another possible condition on permissible interpretations is proportionality. A cause is "proportional" to an effect when it is sufficient for the effect without including unnecessary detail (Yablo, 1992).[15] How exactly to implement proportionality using causal models is an open question. Requiring that the values of a parent variable line up one-to-one with those of the child variable would be extremely difficult to satisfy for interpreted models with more than two variables. A slightly weaker requirement would be that every value of a parent variable be such that intervening to set it to some other value would make a difference to the child variable, where it need not be a unique difference. In the causal model literature, proportionality has been employed as a condition on causal explanation primarily in application to the causal exclusion problem – which challenges the causal efficacy of the mental (Baumgartner, 2009; McDonnell, 2017; Weslake, forthcoming; Woodward, 2008, 2015).[16] There is virtually no attempt to apply it as a condition on the metaphysics of causation.[17]

### §5.3 Stability

Another common condition requires interpreted models be stable – that actual causation verdicts delivered by an interpreted model not flip-flop upon the mere addition or removal of variables (Beckers, 2021; Blanchard & Schaffer, 2017; Fenton-Glynn, 2021; Gallow, 2021; Hall, 2006, 2007; J. Halpern & Hitchcock, 2010; J. Y. Halpern, 2016b).[18] Different variations on a stability condition may concern whether it is the addition, the removal, or both that matter, and whether it is the overturning of a positive verdict, a negative one, or both that matters. However, few see stability as viable on its own. Rather, it should follow from some other, independently-motivated condition. Halpern (2016b), for example, argues that

[15] For a general overview of proportionality, see (Rubenstein, 2023a, 2023b).

[16] For further applications and discussion of proportionality within the SEM framework, see (Franklin-Hall, 2016; Rubenstein, 2024; Woodward, 2010, 2018, 2021).

[17] Though see (McDonald, 2022) for one such attempt.

[18] Of course, a weak kind of stability is ensured by a restriction to accurate models, as previously defined.

stability follows from building a normative parameter into one's analysis, in the way briefly discussed in Part 1, §6.3.

The concern behind stability is that there's something problematic about model relativity – the fact that different interpreted models of the same situation (or set of situations) may deliver different local causal verdicts (Beckers, 2021; Gallow, 2021). A "local causal verdict" is the verdict the recipe delivers about a causal claim relative to a particular interpreted model. This is distinct from a "global causal verdict," which is the verdict the full SEM analysis delivers about this claim, and which takes all apt interpreted models into account by quantifying over them in some way (see Part 1). What exactly is problematic about model relativity is unclear. It seems principally concerning for analyses of causation that universally quantify over a single interpreted model. For them, local causal verdicts just are global ones. Choice of model would be especially difficult to justify, the concern runs, whenever a different causal verdict would have been delivered by another interpreted model, differing from the original by the mere addition of a variable. After all, the new interpreted model is merely an enrichment on the old one. A mere enrichment should not be capable of disrupting the causal relations captured by an interpreted model (Hall, 2006, p. 34). Note that the concern here is over how easily one's SEM analysis of causation can be developed and defended. Without the assurance of stability, an analysis of this form is more difficult to both complete and justify.

This concern is plausibly avoided by any SEM analysis of causation that quantifies over a broader class of apt interpreted models. For them, there is no overturning of a global causal verdict, since disagreement amongst particular interpreted models is factored out of the final result via the quantification.

### §5.3 Addition or Removal of Variables

Concerns over model relativity have also given rise to proposals about when an interpreted model represents *enough* of the situation (Hiddleston, 2005, p. 648). This is somewhat refined into proposals about when a variable may be omitted from an interpreted model, or

when a variable must be added. Gallow (2021), for example, proposes conditions for when a variable can be benignly removed from an interpreted model, and when such removal is impermissible. The proposal is arguably subsumed under one given by McDonald (forthcoming), which goes one step further.

To adopt McDonald's terminology, distinguish a variable that *fully* mediates between its flanking variables from one that only *partially* mediates. A fully mediating variable is one that slots in entirely between flanking variables on either side, replacing one side as parents in the equations for the other. *Y* "fully mediates" between *X* and *Z* just in case *Y* is a child of *X* and a parent of *Z*, and *X* is not a parent of *Z*. Partial mediation, on the other hand, occurs when a variable slots in between its flanking variables without replacing one side as parents of the other. *W* "partially mediates" between *X* and *Z* just in case *W* is a child of *X* and a parent of *Z*, but *X* remains a parent of *Z*.

Gallow proposes that fully mediating variables can be benignly omitted while partially mediating ones cannot. That is, the removal of a partially mediating variable from an apt interpreted model will produce an inapt interpreted model. McDonald argues further that partially mediating variables must be included. That is, an apt interpreted model must include any interpreted variables that would partially mediate between existing interpreted variables were they to be introduced. She calls this "Evident Mediation," and argues that its adoption obviates the need for a normality parameter as a response to the problem of structural isomorphs (see Part 1, §6.3).

## §5 Conclusion

I have been operating under the assumption that a complete and satisfying SEM analysis of general or actual causation requires systematic aptness principles whose application is unambiguous. But perhaps not. It may be that the most we can hope for is a set of heuristics or defeasible principles (Woodward, 2016). Indeed, aptness may be more of an art than a science, essentially requiring creative and independent consideration of the inquiry and target at hand.

This attitude towards aptness has different implications for the nature of causal explanation than for causation itself, though. The more pragmatic nature of causal explanation makes imbuing aptness with a heuristic or artistic character more appropriate. For causation, however, it may undermine the position that causation is mind and language independent or epistemically accessible – that what constitutes causation is something we can identify and track. For a hybrid view (see §3.3), it's worth noting that problems of aptness arise for any other more basic causal notion, as well. Even granting that one or the other concession is ultimately correct, conceding now strikes me as untimely. There is more work still to do. The question of what constitutes an apt interpreted model for the purposes of completing a SEM analysis of causation is significant for precisely the reasons that make it difficult. It calls for clarification around questions of ontology, mereology, modality, and the relationship(s) between them. As such, the challenge of aptness strikes me as remarkably rich and worthy of engagement for reasons that go well beyond the metaphysics of causation.

## §6 References


Andreas, H., & Gunther, M. (forthcominga). A Lewisian Regularity Theory. *Philosophical Studies*.

Andreas, H., & Gunther, M. (forthcomingb). A Regularity Theory of Causation. *Pacific Philosophical Quarterly*.

Baumgartner, M. (2009). Interventionist Causal Exclusion and Non-reductive Physicalism. *International Studies in the Philosophy of Science*, *23*(2), 161–178. https://doi.org/10.1080/02698590903006909

Beckers, S. (2021). Equivalent Causal Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 6202–6209.

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

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.

Casati, R., & Varzi, A. (2000). Topological Essentialism. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *100*(3), 217–236.

Casati, R., & Varzi, A. (2023). Events. *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2023/entries/events/>

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.

Elgin, C. (2004). True Enough. *Philosophical Issues*, *14*(1), 113–131.

Fenton-Glynn, L. (2021). *Causation*. Cambridge University Press.

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. (2021). A Model-Invariant Theory of Causation. *Philosophical Review*.

Gallow, J. D. (2023). Causal Counterfactuals Without Miracles or Backtracking. *Philosophy and Phenomenological Research*, *107*(2), 439–469.

Glynn, L. (2013). Of Miracles and Interventions. *Erkenntnis*, *78*(1), 43–64.

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

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

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

Handfield, T., Oppy, G., Twardy, C. R., & Korb, K. B. (2008). The Metaphysics of Causal Models: Where's the Biff? *Erkenntnis*, *68*(2), 149–168.

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

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. (2003). Of Humean Bondage. *British Journal for the Philosophy of Science*, *54*, 1–25.

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

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

Kroedel, T. (2019). *Mental Causation: A Counterfactual Theory*. Cambridge University Press.

Kuorikoski, J. (2014). How to Be a Humean Interventionist. *Philosophy and Phenomenological Research*, *89*(2), 333–351.

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

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

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

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

Lewis, D. (1986). Events. In D. Lewis (Ed.), *Philosophical Papers Vol. II* (pp. 241–269). Oxford University Press.

Lewis, D. (2000). Causation as Influence. *The Journal of Philosophy*, *97*, 182–197. https://doi.org/10.2307/2678389

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

McDonald, J. (2022). *Actual Causation: Apt Causal Models and Causal Relativism*. The Graduate Center, CUNY. http://academicworks.cuny.edu/gc_etds/4828

McDonald, J. (forthcoming). Essential Structure for Causal Models. *Australasian Journal of Philosophy*.

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

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

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

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

Papineau, D. (2022). The Statistical Nature of Causation. *The Monist*, *105*, 247–275.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (Second edition., 3rd printing..). Cambridge University Press. (Original work published 2000)

Potochnik, A. (2017). *Idealization and the Aims of Science*. Cambridge University Press.

Rubenstein, E. (2023a). Proportionality in Causation, Part 1: Theories. *Philosophy Compass*, *e12957*.

Rubenstein, E. (2023b). Proportionality in Causation, Part II: Applications and Challenges. *Philosophy Compass*, *e12960*.

Rubenstein, E. (2024). Cohesive Proportionality. *Philosophical Studies*, *181*, 179–203.

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. (2012). Causal Contextualism. In M. Blaauw (Ed.), *Contrastivism in Philosophy: New Perspectives*. Routledge.

Starr, W. (2019). Counterfactuals. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2019/entries/counterfactuals/>

Statham, G. (2018). Woodward and Variable Relativity. *Philosophical Studies*, *175*, 885–902.

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. (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. (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.

Woodward, J. (2021). *Causation with a Human Face: Normative Theory and Descriptive Psychology*. Oxford University Press.

Wysocki, T. (2023). An Event Algebra for Causal Counterfactuals. *Philosophical Studies*, *180*, 3533–3565.

Yablo, S. (1992). Mental Causation. *The Philosophical Review*, *101*(2), 245–280. https://doi.org/10.2307/2185535

Yang, E. (2013). Eliminativism, interventionism and the Overdetermination Argument. *Philosophical Studies*, *164*(2), 321–340. https://doi.org/10.1007/s11098-012-9856-0

Zhong, L. (2020). Intervention, Fixation, and Supervenient Causation. *The Journal of Philosophy*, *117*(6), 293–314.