Forthcoming in Philosophy Compass

Causal Models and Metaphysics – Part 1: Using Causal Models[1]

Jenn McDonald

> 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, 2001b, p. 273)

Abstract This paper provides a general introduction to the use of causal models in the metaphysics of causation, specifically structural equation models and directed acyclic graphs. It reviews the formal framework, lays out a method of interpretation capable of representing different underlying metaphysical relations, and describes the use of these models in analyzing causation.

§1 Introduction

Recent work in the philosophy of causation invokes the framework of causal models – namely, structural equation models and directed acyclic graphs. These models come from the special sciences (econometrics, statistics, computer science, etc.), where they have been developed over several decades to understand causal structure and make predictions (Pearl, 2000/2009; Spirtes et al., 1993/2000). A plausible explanation of their success in these endeavors is that they somehow get at the underlying nature of causation. If so, they could shed light on the metaphysics and epistemology of causation. This paper focuses on their role in providing a metaphysical analysis of causation. The focus is significant, as this

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

particular application captures a relatively small portion of the overall use and discussion of causal models.

An analysis of causation in terms of structural equation models (SEMs) and corresponding directed acyclic graphs (DAGs) has at least two components. The first is a definition of causation in terms of a given model or class of models – a "recipe" for reading causal relations off a model. After presenting the formalism and a method of interpretation, the remainder of this paper (Part 1) explores developments and progress with respect to this component. Identifying the right recipe is particularly difficult for actual (i.e., token or singular) causation, which draws the majority attention in the metaphysics literature. In a companion paper (Part 2), I address the second component, which is a principled account of what qualifies a model or class of models as given – or "apt" – such that, in combination with the right recipe, we get a complete causal analysis.

§2 Causal Models

I begin by surveying the basic framework of causal models – structural equation models (SEMs) and directed acyclic graphs (DAGs). A couple notes by way of prelude: first, causal models are versatile – they can represent token- or type-level structures belonging to deterministic or probabilistic systems. Since brevity precludes exhaustive coverage, I restrict discussion to deterministic systems.[2]

Second, many in the literature speak in the same breath of the formal apparatus (aka. "model") and of what it is taken to represent. The utility of this is clear. Aside from the obvious economy of expression, one can simultaneously pull from both formal and real-world causal considerations in developing a formal framework for causal inquiry. However, leaving this distinction implicit risks confusion, and anyway hobbles discussion of a key

[2] For probabilistic models and applications to probabilistic systems, see (Fenton-Glynn, 2021; C. Glymour, 2001; Glynn, 2011; Pearl, 2000/2009; Sloman, 2005; Spirtes et al., 1993/2000; Woodward, 2003), among others.

question: which relationship between a model and its target is of the right kind for the intended project? By distinguishing clearly between the formal model and what it purports to represent, the mapping (aka. "interpretation") that underlies the representation is brought into view. Which representational principles should be met by an apt interpreted model is the principal subject of Part 2.

§2.1 'Nonspecific' Structural Equation Models

On the formalization I assume, structural equation models best suited to represent type-level structure have two components, while those suited to token-level supplement these with an additional component: namely, an assignment.[3] I begin with the first – a "nonspecific" SEM.

A nonspecific SEM is an ordered pair, $\mathcal{M}_i = < \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{L}} >$, built of a signature and a linkage. The "signature" is a collection of variables. It includes a set of exogenous variables (roughly, the independent variables), a set of endogenous variables (the dependent ones), 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 values to each variable, $X: X \in (\boldsymbol{U} \cup \boldsymbol{V})$.[4] Values of variables represent causal relata. For example, a variable with two values can represent an action being taken or not, and a variable with many values can represent the mass of a particular object. The next section explores this further.

The second component – the *linkage* – is a set of asymmetric functional equations defined over the signature. Each equation, $X \coloneqq f(\boldsymbol{PA}(X))$, indicates an endogenous variable on the left-hand side, X, and a function on the right-hand side over a subset of remaining variables

[3] So as to focus on philosophical applications, I survey just one of various formalizations of SEMs in the literature, which mostly follows Halpern (2000); but see also (Blanchard & Schaffer, 2017; Gallow, 2023).

[4] A word on notation: an uppercase bolded letter ('$\boldsymbol{U}$') indicates a set, an uppercase italicized letter ('X') indicates a variable, a lowercase italicized letter ('x') indicates a value of a variable, and a vector symbol ('$\overrightarrow{\boldsymbol{X}}$') over the name of a set indicates an ordered set.

from the signature, $\boldsymbol{PA_X} \subseteq (\boldsymbol{U} \cup \boldsymbol{V})\backslash X$. This specifies the functional relationship between the "child" variable on the left-hand side, *X*, and the "parent" variables on the right-hand side, ***PA***. The function takes in the values of the parent variables and puts out a value of the child variable, and can be given in many forms. For example, it might utilize arithmetic ('$Y \coloneqq \frac{3}{2}X + Z$') or Boolean algebra ('$Y \coloneqq X \vee W$') – whatever best captures the worldly dependence relations that the equations are taken to represent.

Crucially, these equations are *asymmetric*. Each says something about the results of any intervention whose target is a parent variable, but says nothing about an intervention whose target is the child. An "intervention" is a mathematical operation on a variable in a model that forces that variable to one of its values, decoupling it from its parent variables, and leaving all other equations untouched. More precisely, an intervention on a model, $\mathcal{M}$, targets a set of variables, $\vec{\boldsymbol{X}}$, setting each variable, $X_i \in \vec{\boldsymbol{X}}$, to one of its values, $x_j \in \boldsymbol{R}(X_i)$. This produces a sub-model, $\mathcal{M}_{\vec{\boldsymbol{X}}=\vec{\boldsymbol{x}}}$, which is identical to $\mathcal{M}$ save that each X_i-equation – the equation with 'X_i' on the left-hand side – is replaced with the equation $X_i = x_j$.[5] Since an intervention on a child variable decouples it from its parents, by definition an intervention on the child variable is silent on the values of the parent variables. But an intervention on a parent variable results in the child variable taking whatever value is put out by the specified function over its parent variables in light of the intervention. In fact, this condition is definitive of the parenthood relation. A variable, *X*, is a "parent" of *Y* if and only if there is an intervention on *X* that, while holding fixed by intervention all other variables aside from *Y* at some one of their values, would result in a change in the value of *Y*. Formally, *X* is a "parent" of *Y* if and only if for some two values of *X*, x_1 and x_2, where $x_1 \neq x_2$, and some two values of *Y*, y_1 and y_2, where $y_1 \neq y_2$, an intervention on *X* to change it from $X = x_1$ to $X = x_2$, while holding each of the remaining variables, $\vec{\boldsymbol{Z}} = (\boldsymbol{U} \cup \boldsymbol{V}) \setminus X, Y$, fixed at some assignment of values, $\vec{\boldsymbol{z}}$: $z_i \in \boldsymbol{R}(Z_i)$, results in a change in *Y* from $Y = y_1$ to $Y = y_2$. Note the existential quantifiers – all that's required for parenthood is that there be one qualifying pair of values of *X*, one pair of

[5] This follows Pearl (2000/2009, sec. 3.2.1), see also (Briggs, 2012; J. Y. Halpern, 2016). For a different implementation, see (Fenton-Glynn, 2021; Pearl, 1993, 2000/2009, sec. 3.2.2; Woodward, 2003).

values of Y, and one assignment to the remaining variables. Equations are *minimal* in that the right-hand side includes only parent variables.

It is assumed that the semantics for any counterfactual framed purely in terms of a model (*e.g.,* 'Had $X = x_2$, then $Y = y_3$'), treats the antecedent as set by intervention. The counterfactual is true, then, only if it's consequent holds in the sub-model produced by this intervention. So far as I have defined it, an intervention is a purely formal operation. I discuss what it might represent in Part 2.

Finally, the metaphysics literature focuses on recursive SEMs. "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 rules out cycles.[6]

§2.2 'Specific' Structural Equation Models

To represent structure at the token-level, a SEM includes a specification of value for each exogenous variable, $X \in \boldsymbol{U}$. This is either treated as extraneous to the model – as an "interpretation" or "context," or it is incorporated into the model as a third component – as an "assignment." I follow the latter, in part to reserve "interpretation" for reference to the assignment of real-world content to variables (see §3.1). On this framework, a "specific" SEM is an ordered triple $\mathcal{M}_i = < \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{L}}, \boldsymbol{\mathcal{A}} >$, with a signature, a linkage, and an assignment. The *assignment* is a function, $\boldsymbol{\mathcal{A}}$, that, to every variable $U\colon U \in \boldsymbol{U}$, maps a value $u\colon u \in \boldsymbol{R}(U)$. Representing each mapping as a constant equation, each included in the model's complete set of structural equations, streamlines the formalism by permitting interventions on exogenous variables. Note that this treatment is not typical. Recursivity, coupled with a requirement that there is only one equation for each variable, entails that endogenous variables receive a unique value under any assignment as determined by the linkage.

[6] This simplifies the dialectic, but rules out cycles by fiat. Ideally, recursive models will follow from aptness principles – such as accuracy – rather than needing to be assumed. In general, non-recursive models can coherently represent causal structures.

§2.3 Directed Acyclic Graphs

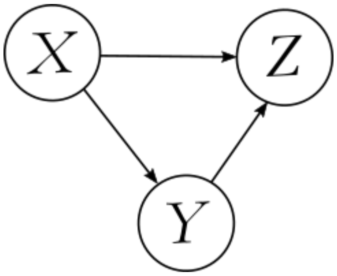
Recursive SEMs can be represented by directed acyclic graphs (DAGs). A DAG consists of nodes with directed edges, or arrows, connecting them. Arrows are drawn from parent variables to child variables. A DAG represents the qualitative information contained in a SEM and serve as a useful visual aid.[7] Here is a sample specific SEM and corresponding DAG:

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

Assignment:
(EQ1) $X = 1$

Linkage:
(EQ2) $Y := 1 - X$
(EQ3) $Z := \max(X, Y)$

Directed Acyclic Graph:



$\mathcal{M}_1$

§3 A Proposed Method of Interpretation

A causal model needs to be interpreted to say something about the world. There is no agreed upon method of interpretation in the literature. But there is enough overlap to invite systematization. I here focus on a basic method and on non-controversial principles of interpretation. Greater detail and controversy come in Part 2.

§3.1 Defining "Interpretation"

[7] Note that DAGs serve many other useful functions in the broader causal model literature. For example, under certain assumptions, one can discern correlational (in)dependencies from a DAG.

Call an "interpretation" of a SEM an assignment of real-world content to its variables. An "interpretation" maps each value of each variable onto an actual or possible factor. A *factor* is anything reasonably treated as causal relata – events, properties, property instances, propositions, etc. The causal model framework seems, and the literature generally is, neutral on the nature of causal relata. (Part 2 shows how the neutrality only goes so far.) Any commitments regarding causal relata simply generate aptness principles on interpreted models. A commitment to relata as property instances, for example, generates an aptness principle whereby the interpretation map only property instances to values of variables. Not to commit, but to simplify discussion, I assume relata are property instances – i.e., a particular object instantiating a specified property over a particular time period.

Given this, an interpretation will assign to each variable a (possibly singleton) set of objects and to each of the variable's values a property possibly instantiated by each object in that set.[8] One way to formalize this is to treat an interpretation as assigning a range of properties to each variable, on the one hand, and a population to the SEM, on the other, where the "population" is a set of n-tuples of objects. Each n-tuple maps one-to-one onto the variables of the SEM, (so, $n = |\{\boldsymbol{U} \cup \boldsymbol{V}\}$), with the i^{th} member of each n-tuple mapping to the same variable. Of course, the same object may appear more than once in a given n-tuple. There are two kinds of interpretation: general and particular. A "general" interpretation assigns a population of at least two members. It can therefore be used to capture how many different situations each could go (or have gone) differently. A "particular" interpretation assigns a singleton population. It captures various ways a single situation could go (or have gone).

§3.2 "Permissible" Interpretations

Even assuming the above, not just any assignment of content will do. A "permissible" interpretation satisfies, at least, exclusivity, exhaustivity, and distinctness. "Exclusivity" requires that whatever properties are mapped onto any two values of a given variable be

[8] While some explicitly incorporate it (Beckers & Vennekens, 2018), the time parameter is almost universally left implicit. I follow suit, leaving the time period largely implicit.

mutually exclusive. The door being closed will be mutually exclusive with the door being open (at the same time). Exclusivity ensures that, in representing the world, a variable will never take more than one value at a time.[9] "Exhaustivity" requires that the set of all properties mapped onto the full range of values of a given variable be jointly exhaustive. The child being under or equal to the height of 36 inches or over 36 inches will be exhaustive. This ensures that a variable will always take at least one value.[10] Exclusivity and exhaustivity are purely formal requirements on the causal model framework, necessitated by its inability to accommodate a variable simultaneously taking more than one value or one failing to take a value – which, to be clear, is a distinct matter from a variable taking the value '0'. What exactly *counts* as mutually exclusive or jointly exhaustive, on the other hand, calls up a host of difficult representational questions. These will be addressed in Part 2.

"Distinctness" is needed for theoretical reasons. It requires that whatever property instances are mapped onto any two values of different variables be "distinct" on some to-be-defined notion.[11] Roughly, different variables must represent property instances that are logically, conceptually, and otherwise metaphysically independent from each other. This ensures that an intervention on a model will never represent something metaphysically impossible. It also serves the same purpose as it did for a traditional counterfactual analysis – it separates the wheat of causation from the chaff of mere counterfactual dependence.

The need for exclusivity, exhaustivity, and distinctness is universally acknowledged. What exactly *counts* as mutually exclusive, jointly exhaustive, or distinct, on the other hand, remains unsettled. Again, this will be taken up in Part 2, alongside other permissibility conditions to consider.

[9] See (Blanchard & Schaffer, 2017, p. 182; Briggs, 2012, p. 142; Hitchcock, 2004, p. 145, 2007b, p. 76, 2007a, p. 502; Pearl, 2000/2009, p. 3; Woodward, 2003, p. 98).

[10] See (Blanchard & Schaffer, 2017, p. 182; Briggs, 2012, p. 142; Hitchcock, 2001b, p. 287; Pearl, 2000/2009, p. 3; Woodward, 2016, p. 1064).

[11] Distinctness is initially proposed and developed by Lewis in a pure counterfactual context. See especially (Lewis, 1986). For references in a SEM context, see (Blanchard & Schaffer, 2017, p. 182; Briggs, 2012, p. 142; Hitchcock, 2004, p. 146, 2007a, p. 502; Paul & Hall, 2013, p. 59).

§3.3 What Equations Represent

What else an interpreted model says depends on one's view about the ontological nature of the relations represented by the equations. The principal positions are, effectively, the traditional ones: regularities or counterfactuals (leaving open which semantics is relevant). Full discussion comes in Part 2.[12]

§4 The Common Form of a SEM Analysis of Causation

Now that we have interpreted models, the question is how they might be used to analyze causation. It is convenient to treat any causal model analysis of causation as having the same form. I mentioned earlier two components to any such analysis: a recipe and an account of aptness. A recipe is a set of conditions both necessary and sufficient for a given variable in the target model to count as a 'type-level cause' of another variable, or for a given value of a variable to count as an 'actual cause' of some value of another variable. To then procure causal verdicts about the world, the model needs to be interpreted. To get the right causal verdicts, though, we need to say what counts as an apt interpreted model. This is the focus of Part 2. Here, we can recognize that a problem will arise whenever two apt interpreted models deliver conflicting verdicts about what causes what. To handle this, a SEM analysis quantifies in some way over the class of apt interpreted models. Thus, any analysis also has a third component: a quantifier selection.

Many analyses employ the existential quantifier: *c* is a cause of *e* just in case there is *at least one* apt interpreted model satisfying the recipe (Blanchard & Schaffer, 2017; Hitchcock, 2001b, 2007a; Weslake, 2015, forthcoming; Woodward, 2008). But any quantifier is logically permitted. Others, for example, employ a universal quantifier (Hall, 2007). While not exactly trivial, the choice of quantifier is less substantive than it may first appear. It impacts what

[12] The SEM framework is surprisingly neutral here, as well. This suggests that a precise grasp of the underlying metaphysics just isn't needed for causal inference.

needs to be said about aptness, and so pairs off with an account of aptness. An existential quantifier requires ruling out models which mistakenly witness causation where there isn't any. A universal quantifier requires ruling out models which mistakenly witness the absence of causation where there really is. Some analyses, however, talk not of quantifying over apt interpreted models but of *justifying* their choice of interpreted model (Beckers & Vennekens, 2018; J. Y. Halpern, 2016; J. Y. Halpern & Hitchcock, 2015; J. Y. Halpern & Pearl, 2005). I see these as simply employing a universal quantifier with a stricter account of aptness – one that permits as apt only a single interpreted model, or perhaps a single equivalence class of them (Beckers, 2021).

Any SEM analysis therefore has the following form:

> **Causation – << *K* >>** $_{\mathbf{SEM}}$ *c* is a << *insert kind of causal relation (K)* >> of *e* just in case << *insert quantifier* >> interpreted models that satisfy << *insert aptness principles* >> and according to which < *c, e* > satisfies << *insert recipe* >>.

§5 Causation at the Type-Level

Traditionally, the complement to actual causation (to be discussed shortly) is taken to be "general" causation – a repeatable causal relation holding between types. General causation is reflected in claims like:

> Erupting volcanos cause ash clouds to form.
> Truancy causes poor performance in school.
> Exposing fragile objects to force cause them to break.

But discussion of general causation in the SEM framework is something of a lacuna (as a metaphysical relation, at least, rather than a kind of causal explanation). Instead, attention is paid to relations that hold between what can be represented by variables. Call this "causal influence."[13] A variable represents a range of properties instantiable by a situation or set thereof. Either way, relations represented as holding between variables are not *particular* relations. Causal influence, then, is the natural complement to actual causation in the SEM framework. An analysis of general causation could be given, in the common form laid out above, that invokes causal influence in its recipe. But the SEM framework permits articulation of several different ways two variables might be related. Which to invoke? What should the recipe otherwise look like? There is no extant proposal. Instead, the SEM literature focuses on identifying different causal influence relations and considering their applications. A brief survey of these relations follows.

§5.1 Causal Relevance and Causal Influence

A variable can be causally relevant to another in several ways:[14] X can be a *direct cause*, a *total cause*, and/or a *contributing cause* of Y.[15] When the variables are given a general interpretation, the represented relations resemble general causation. When given a particular interpretation, they resemble contrastive token-level causal relations.

Direct causation holds just in case an intervention on one variable leads to a change in another when all others are held fixed by intervention. So, a "direct cause" is just a parent variable. But notice that X will no longer count as a direct cause of Y if a variable, W, is fully inserted between them (replacing X in the Y-equation). Since W is fixed when establishing parenthood, any dependence of Y on X is screened off.

[13] The name is inspired by, but the notion is not identical to, that in Lewis (2000).

[14] This presentation follows Woodward (2003), departing only to the extent that he, like most others, uses 'variable' loosely to refer simultaneously to formal items in a model as well as to the worldly properties these items represent. As a result, his account of causal relevance between variables simultaneously covers both formal relations between variables and worldly relations between sets of property instances. This, of course, requires an implicit assumption in the background that the models are apt.

[15] These definitions are developed in part from those offered by Pearl (2000/2009). See also (Hitchcock, 2001a) for the distinction between total and contributing causation in different terms.

Total causation holds when there is some assignment to the exogenous variables such that an intervention on one variable changes another (holding nothing else fixed). Formally, X is a "total cause" of Y just in case for some two distinct values of X, x_1 and x_2, and some two distinct values of Y, y_1 and y_2, an intervention from $X = x_1$ to $X = x_2$ results in a change from $Y = y_1$ to $Y = y_2$. While broader in some ways (by capturing many ancestors), total causation is also narrower than direct causation. X from $\mathcal{M}_1$ is a direct cause, but not a total cause, of Z. Additionally, both direct and total causation are silent on the relation between two variables when the influence of one sort (positive, say) exactly cancels out the influence of another (negative, say).[16]

Contributing causation captures nuanced relations like these.[17] It holds when there is a *chain* of variables between X and Y such that a change in X leads to a change in Y when variables not on that chain are fixed. Formally, define a "directed route," $\overrightarrow{\boldsymbol{R_\iota}}$, between X and Y, hereafter a "route", as a sequence of variables, $\overrightarrow{\boldsymbol{R_\iota}} = < X, W_1, \dots, W_i, Y >$, such that X is a parent of W_1, …, W_{i-1} is a parent of W_i, and W_i is a parent of Y. (The sequence of corresponding nodes in a DAG will be such that the arrows between them all point in the same direction.) For a given route, $\overrightarrow{\boldsymbol{R_\iota}}$, between X and Y, any member of $\overrightarrow{\boldsymbol{R_\iota}}$, including X and Y, is an "on-route" variable, while all other variables in the model, $Z \in \boldsymbol{Z} = \{(\boldsymbol{U} \cup \boldsymbol{V})\backslash \boldsymbol{R_i}\}$, is an "off-route" variable. Then, X is a "contributing cause" of Y just in case there is a route, $\overrightarrow{\boldsymbol{R_1}}$, between X and Y and a setting of off-route variables, $\vec{\boldsymbol{Z}} = \vec{\boldsymbol{z}}$: $z_i \in \boldsymbol{R}(Z_i)$, such that for some two distinct values of X, x_1 and x_2, and some two distinct values of Y, y_1 and y_2, an intervention to change X from $X = x_1$ to $X = x_2$, while $\vec{\boldsymbol{Z}} = \vec{\boldsymbol{z}}$ is held fixed, results in a change in Y from $Y = y_1$ to $Y = y_2$.

§6 Actual Causation

[16] Called 'failures of faithfulness'. See, for example, (Spirtes et al., 1993/2000).

[17] Though it will leave out certain ancestors due to what Hitchcock calls a "failure of composition." This occurs, for example, when a simple chain, $X \rightarrow Y \rightarrow Z$, is such that while an intervention on X affects Y and an intervention on Y affects Z, no intervention on X alone affects Z. See especially Hitchcock's "Dog Bite" (2001b, pp. 290–291).

Moving on to the vexed task of analyzing *actual* causation. Here, the SEM framework seems to have made for recent progress. Actual causation, aka. "token" or "singular" causation, holds between two particular property instances when the first causes the second. This is reflected in claims like:

> 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.

As discussed above, a SEM analysis provides necessary and sufficient conditions for when an actual causal relation holds between two particular things in terms of properties of an apt interpreted model or class of them. Different proposals vary along two dimensions. First, in what constitutes the relevant properties of a model, or "recipe." Second, and largely independently, in the underlying metaphysical project – specifically in reductive aspirations. In general, a SEM analysis reduces actual causation to whatever the equations are taken to represent. But this can be different things. If the equations represent complex counterfactuals, then actual causation reduces to counterfactual dependence. If, instead, the equations are interpreted as representing causal influence, then actual causation reduces to causal influence. (Then, if causal influence reduces further, actual causation does too.) Interestingly, interlocutors of either camp debate the former dimension – the recipe, that is – seemingly unimpeded by divergent metaphysical projects. This is, arguably, one of the more interesting contributions of causal models. It allows the discussion to progress in purely formal terms, without having to first resolve underlying metaphysical disagreements. The remainder of this section addresses the current state of that debate, relegating metaphysics to Part 2.

§6.1 A Simple Recipe and Redundant Causation

A simple recipe, to start, is that *c* is an actual cause of *e* relative to an interpreted model just in case intervening to change the value of the variable that represents *c* leads to a change in the value of the variable that represents *e*. The eruption of Mount Vesuvius in 79 AD is an actual cause of the city of Pompei being buried in ash in 79 AD relative to an interpreted model just in case intervening on the eruption variable to change its value from 'erupt' to some other value leads to a change in the value of the buried-in-ash variable from 'buried' to some one of its other values.

The problem is that this won't cover cases of redundant causation, when there are at least two different property instances serving as causes of the same effect, either of which is sufficient. In overdetermination cases, both property instances are (intuitively) causes. In preemption cases, only one is a cause while the other stands in the wings, so to speak, ready to step in should the first fail. Take, as an example, a case of late preemption:

> **Early Preemption** Suzy throws a rock at a window, which shatters. Billy stands by, but would have thrown instead had Suzy failed to. Suzy and Billy have equal strength and excellent aim, and their rocks are of equal weight.

The intuition is that Suzy's throw causes the window to shatter. But the simple recipe doesn't deliver this result. The interpreted model, $< \mathcal{M}_1,\ \mathcal{I}(\mathcal{M}_1) >$, from before is plausibly apt, but it doesn't deem Suzy's throw a cause. An intervention changing X from 1 to 0 doesn't result in a change in Z. Regardless of whether the aim is a conceptual analysis, functional analysis, or providing a real definition,[18] the accommodation of vivid intuitions like that above is at least a defeasible desideratum on a satisfactory analysis. [19]

§6.2 De Facto Recipes: Holding Fixed "Off-Route" Variables

[18] A "real definition" of *x* is an account of what *x* is in the world, as opposed to a semantic account of the term 'the *x*' or a conceptual analysis of our concept of *x* (Rosen, 2015).

[19] For conceptual or functional analyses, this results from the principal target being a *normative* analysis. By this, I mean an "explication" in Carnap's sense (1962, 1988) – a refinement on a typically ambiguous or opaque concept that involves revision, thus resulting in a notion that will likely fail to accord with all intuitions or pre-theoretic uses of said concept.

The principal response in the SEM literature is motivated by the idea that dependence between the effect and the genuine cause can be uncovered once certain factors in the situation are held fixed either as they actually occur or as they might have occurred. The dependence between the window shattering and Suzy's throw is uncovered if we hold fixed the fact that Billy doesn't throw. This is the idea of "de facto" dependence (Yablo, 2002, 2004). To implement this idea in the SEM framework, we invoke the above distinction between "on-route" and "off-route" variables in the following schema.[20]

Recipe – AC $_{\textbf{Schema}}$ $X = x$ is an actual cause of $Y = y$ in $\mathcal{M}_i$ just in case…

[1] $X = x$ and $Y = y$ in $\mathcal{M}_i$.

[2] There is a directed route R_i in $\mathcal{M}_i$ from X to Y and, for the set of variables off R_i, $\vec{Z}$, a permissible assignment of values, $\vec{z}$, such that:

(a) Had $\vec{Z} = \vec{z}$ and $X = x$, then $Y = y$.

(b) Had $\vec{Z} = \vec{z}$ and $X = x_i$, where $x_i \neq x$, then $Y = y_i$, where $y_i \neq y$.[21]

[1] is the actuality condition. When coupled with the model's aptness, which requires that the model say only true things, this ensures that the cause (which would be represented by $X = x$) and the effect (represented by $Y = y$) do actually occur.

[2] is the de facto causal condition. It says that there must be a route between the putative cause variable, X, and the effect variable, Y, such that when all off-route variables are held fixed at values that satisfy some to-be-defined permissibility condition, then intervening to set the putative cause as occurring ($X = x$) will result in the effect occurring ($Y = y$), and

[20] Some recipes implement this idea in terms of sets of variables, rather than routes – replacing 'directed routes' with 'sets' and 'off-route variables' with 'complement set' (J. Y. Halpern, 2016; J. Y. Halpern & Pearl, 2005). Note that such recipes will be roughly equivalent, since a route is simply an ordered set, with off-route variables constituting the complement set.

[21] Pearl (Pearl, 2000/2009, Chapter 10) makes the first proposal of this kind, though not in this exact form.

intervening to set some alternative to the putative cause ($X = x_i$) will result in some alternative to the effect ($Y = y_i$).

§6.2 Permissibility of "Off-Route" Variable Assignments

The open question is what assignments of values to off-route variables count as permissible. What contingencies, exactly, are relevant to causation? An initial proposal is that only actual values be permitted:

P-0 $\vec{Z} = \vec{z}$ are the values given by $\boldsymbol{\mathcal{A}}_{\mathcal{M}_1}$ and $\boldsymbol{\mathcal{L}}_{\mathcal{M}_1}$.

This handles early preemption (Billy doesn't actually throw), and late preemption, as well. To illustrate, consider:

Late Preemption Suzy and Billy are throwing rocks at a window. Suzy throws just before Billy. Both rocks fly straight to the window, with Suzy's rock hitting first and shattering it. Billy's rock flies moments later through the empty space the window had just occupied.

The relevant actual factor to hold fixed is that Billy's rock doesn't hit the window. (After all, there's no window to hit!) Then, the dependence between the window shattering and Suzy's throw is revealed.[22] However, **P-0** doesn't handle overdetermination cases. Consider:

Overdetermination Suzy and Billy both throw rocks, which hit the window simultaneously and it shatters. Either rock is of sufficient mass and hits with sufficient force that the window would have shattered had it been hit only with one or the other.

The following proposal responds:

[22] However, see (Hall, 2007) for reason to think there's something amiss in this solution.

P-1 Had $\vec{Z} = \vec{z}$, then $\overrightarrow{R_\imath} = \overrightarrow{r_\imath}$, where $\overrightarrow{r_\imath}$ is the set of values given by $\boldsymbol{\mathcal{A}}_{\mathcal{M}_1}$ and $\boldsymbol{\mathcal{L}}_{\mathcal{M}_1}$.[23]

P-1 requires that the setting of off-route variables preserve the actual values of the on-route variables. This is strictly more permissive, in that anything permitted by **P-0** will also be permitted by **P-1**. Since a token-level SEM has a unique solution for any assignment, the actual values of $\overrightarrow{R_\imath}$ are obviously consistent with the actual values of $\vec{Z}$.

P-1 allows consideration of the dependence of the window shattering on Suzy's throw (or Billy's), under the contingency that Billy (or Suzy) doesn't throw. Thus revealing Suzy's throw (or Billy's) as an actual cause.

§6.3 Further Conditions?

However, cases can be constructed wherein **P-1** fails to deliver the right results. In response, other conditions have been proposed, either to replace or supplement. A difference-making condition, for example, requires that no alternative to the actual cause, had it occurred instead, would have qualified as an actual cause of the *same* effect (Beckers & Vennekens, 2017, 2018; Weslake, 2015).[24]

Alternatively, many have proposed incorporating a distinction between default values of a variable (i.e., normal or typical values) and deviant ones (i.e., abnormal or atypical ones) (Andreas & Gunther, forthcoming; Gallow, 2021; Hall, 2007; J. Y. Halpern, 2008; J. Y. Halpern & Hitchcock, 2013, 2015; Hitchcock, 2007a; Livengood, 2013; Menzies, 2004a, 2004b; Paul & Hall, 2013). This is principally (though not always[25]) in response to the *problem of structural isomorphs*. 'Structural isomorphs' are different situations that can be aptly represented by the same model, under different interpretations. The problem is that the

[23] Variations on **P-1** can be found in (J. Y. Halpern, 2016; J. Y. Halpern & Pearl, 2005; Hitchcock, 2001b; Woodward, 2003, pp. 83–84).

[24] Sartorio (2005) proposes this condition, as well, for a traditional (i.e. non-SEM) counterfactual analysis.

[25] Considerations regarding voting scenarios have also generated a normality parameter.

same model can be interpreted so as to aptly represent two situations with intuitively different causal structures. Insofar as a given recipe gets one situation right, it ipso facto gets the other wrong.

Proponents of a normality parameter point out that in such cases, the normative role played by the intuitive cause in the one situation and analogous non-cause in the other are different. They argue that this is what explains the difference in intuition, and further argue that this justifies incorporating relativity to norms into our analysis of actual causation. This incorporation may be implemented by doing three things: (1) supplementing the formalism with a tag for each value of each variable indicating whether it is default or deviant (perhaps with a formal requirement that only one value of each variable be tagged default; or perhaps permitting weighted measures of normality);[26] (2) adding conditions requiring that the normality claims made by a permissible interpretation be true (perhaps making explicit they be true relative to the same set of norms); and (3) adjusting the recipe so that a variable's value only counts as a cause if intervening to change it to a *more normal* value results in the relevant change in the effect variable.

Resistance to this proposal takes various forms: that incorporating a normality parameter is untenable and psychologically implausible (Blanchard & Schaffer, 2017), that the right aptness principles adequately respond to the problem of structural isomorphs (Blanchard & Schaffer, 2017; McDonald, forthcoming), and/or that a normality parameter fails, after all, to so respond (Wysocki, forthcoming).

§7 Conclusion

[26] Alternatively, the normality ordering might be set over an assignment of values to exogenous variables (J. Y. Halpern, 2016), or over complete assignments of values to variables (J. Y. Halpern, 2008; J. Y. Halpern & Hitchcock, 2015), though see (J. Y. Halpern & Hitchcock, 2013) for a translation between this and an ordering over individual variables.

The SEM framework has galvanized the philosophy literature on causation, though many questions remain. In addition, applications of the SEM framework explore beyond the nature of causation, asking after its character. For example, whether causation is transitive (Beckers & Vennekens, 2017; Hitchcock, 2001b), and whether there is a substantive difference between causing and preventing (Kroedel, 2019; Paul & Hall, 2013). Causal models have also been used to calculate the cardinality of kinds of causal structures, which far exceeds the number of those closely examined in the literature (C. N. Glymour et al., 2010). This poses a challenge to the method of cases in evaluating analyses of causation.[27] However, such methodology arguably retains merit (Paul & Hall, 2013, Chapter 6).

§8 References

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

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

Beckers, S., & Vennekens, J. (2017). The Transitivity and Asymmetry of Actual Causation. *Ergo*, *4*(1), 1–27.

Beckers, S., & Vennekens, J. (2018). A Principled Approach to Defining Actual Causation. *Synthese*, *195*(2), 835–862.

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

[27] Also called "method by counterexample" (Paul & Hall, 2013).

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

Carnap, R. (1962). *Logical Foundations of Probability* (Vol. 2). University of Chicago Press.

Carnap, R. (1988). *Meaning and Necessity: A Study in Semantics and Modal Logic*. University of Chicago Press.

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

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.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT Press.

Glymour, C. N., Danks, D., Eberhardt, F., Ramsey, J., Scheines, R., Spirtes, P., Teng, C. M., & Zhang, J. (2010). Actual Causation: A Stone Soup Essay. *Synthese*, *175*, 169–192.

Glynn, L. (2011). A Probabilistic Analysis of Causation. *The British Journal for the Philosophy of Science*, *62*(2), 343–392.

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

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

Halpern, J. Y. (2008). Defaults and Normality in Causal Structures. *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, 198–208.

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

Halpern, J. Y., & Hitchcock, C. (2013). Compact Representations of Extended Causal Models. *Cognitive Science*, *37*, 986–1010.

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

Halpern, J. Y., & 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

Hitchcock, C. (2001a). A Tale of Two Effects. *Philosophical Review*, *110*, 361–396.

Hitchcock, C. (2001b). 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.

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

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

Livengood, J. (2013). Actual Causation and Simple Voting Scenarios. *Nous*, *47*(2), 316–345.

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

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.

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

Pearl, J. (1993). Comment: Graphical Models, Causality, and Intervention. *Statistical Science*, *8*(3), 266–269.

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

Rosen, G. (2015). Real Definition. *Analytic Philosophy*, *56*(3), 189–209.

Sartorio, C. (2005). Causes as Difference-Makers. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *123*(1/2), 71–96.

Sloman, S. (2005). *Causal Models: How People Think About the World and its Alternatives*. Oxford University Press.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. Springer-Verlag. (Original work published 1993)

Weslake, B. (2015). A Partial Theory of Actual Causation. *Manuscript*.

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). Response to Strevens. *Philosophy and Phenomenological Research*, *LXXVII*(1), 193–212.

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072. https://doi.org/10.1007/s11229-015-0810-5

Wysocki, T. (forthcoming). Conjoined Cases. *Synthese*.

Yablo, S. (2002). De Facto Dependence. *The Journal of Philosophy*, *99*(3), 130–148.

Yablo, S. (2004). Advertisement for a Sketch of an Outline of a Proto-Theory of Causation. In N. Hall, L. A. Paul, & J. Collins (Eds.), *Causation and Counterfactuals* (pp. 119–137). MIT Press.