

Interventionist counterfactuals

Rachael Briggs

Published online: 23 May 2012
© Springer Science+Business Media B.V. 2012

Abstract A number of recent authors (Galles and Pearl, *Found Sci* 3 (1):151–182, 1998; Hiddleston, *Noûs* 39 (4):232–257, 2005; Halpern, *J Artif Intell Res* 12:317–337, 2000) advocate a causal modeling semantics for counterfactuals. But the precise logical significance of the causal modeling semantics remains murky. Particularly important, yet particularly under-explored, is its relationship to the similarity-based semantics for counterfactuals developed by Lewis (Counterfactuals. Harvard University Press, 1973b). The causal modeling semantics is both an account of the *truth conditions* of counterfactuals, and an account of which inferences involving counterfactuals are *valid*. As an account of truth conditions, it is incomplete. While Lewis's similarity semantics lets us evaluate counterfactuals with arbitrarily complex antecedents and consequents, the causal modeling semantics makes it hard to ascertain the truth conditions of all but a highly restricted class of counterfactuals. I explain how to extend the causal modeling language to encompass a wider range of sentences, and provide a sound and complete axiomatization for the extended language. Extending the truth conditions for counterfactuals has serious consequences concerning valid inference. The extended language is unlike any logic of Lewis's: *modus ponens* is invalid, and classical logical equivalents cannot be freely substituted in the antecedents of conditionals.

Keywords Counterfactuals · Causal modeling · *Modus ponens* · Simplification of disjunctive antecedents

Earlier versions of this paper were presented at the 2011 Carolina Metaphysics Workshop, the 2010 Formal Epistemology Workshop, and various workshops and departmental seminars around Australia.

R. Briggs (✉)
Australian National University, Canberra, ACT, Australia
e-mail: formal.epistemology@gmail.com

R. Briggs
Griffith University, Nathan, QLD, Australia

1 Introduction

Causal models are invaluable to statisticians, who aim to understand and predict complex statistical relationships between variables. Philosophers have adopted causal models as a tool for analyzing causation and causal explanation (Handfield et al. 2008; Woodward 2003; Menzies 2004; Hitchcock 2001).

Since counterfactuals are closely connected to causation, it would be worthwhile to develop a causal modeling semantics for counterfactuals. Such a semantics should answer two questions. The first of these questions is about *truth conditions*: which features of the world determine the truth values of counterfactuals, and how do they determine the truth values of counterfactuals? The second question is about *validity*: which inferences containing counterfactuals are valid? In answering both questions, it will be useful to compare causal modeling accounts with the older similarity-based account of Lewis (1973b).

Galles and Pearl (1998) have proposed a partial account of the truth conditions of counterfactuals; similar accounts can be found in (Hiddleston 2005) and (Halpern 2000). Woodward (2003, p. 137) argues that the causal modeling account disagrees with Lewis's account in one crucial respect: it assigns different truth conditions to counterfactuals. Galles and Pearl (1998) argue that the causal modeling account agrees with Lewis's account in another crucial respect, however: it counts the same inferences as valid.

I will argue that the divergence between the two accounts is larger than commonly recognized: they disagree both about the truth conditions of counterfactuals, and about which inferences containing counterfactuals are valid. The causal modeling account, as developed so far, assigns truth conditions to a narrower range of counterfactuals than the similarity account. We can extend the causal modeling account to assign truth conditions to a wider range of counterfactuals. Extending the semantics, however, has drastic logical consequences.

My basic plan is as follows. In the remainder of the introduction, I will outline a general selection-based framework, and explain the similarity and causal modeling accounts as different ways of cashing out this framework. In Sect. 2, I will compare and contrast the similarity and causal modeling accounts on the subjects of truth conditions and validity. The causal modeling account turns out to have a crucial limitation: it provides truth conditions for only a narrow range of counterfactuals. In Sect. 3, I will overcome this limitation, and extend the causal modeling semantics. In Sect. 4, I will give a sound and complete axiomatization of the extended causal modeling language, and discuss the relationship between the extended causal modeling logic and the logic of similarity-based accounts.

1.1 A common framework

Both the similarity semantics and the causal modeling semantics are special cases of the same general framework, which relies on the concept of a *selection function* (proposed by Stalnaker (1981b) and refined by Lewis (1973b)). The framework assigns truth values to counterfactual conditionals—sentences of the form “If ϕ had been the case, then ψ would be the case”, abbreviated “ $\phi \Box \rightarrow \psi$ ”—at *points*.

The similarity semantics interprets the points as possible worlds; the causal modeling semantics interprets them as models.

A selection function maps each point α and antecedent ϕ to a set of points $f(\phi, \alpha)$. $\phi \Box \rightarrow \psi$ is true at α if ψ is true at every point in $f(\phi, \alpha)$ and false otherwise.¹

The similarity semantics and the causal modeling semantics place different interpretations on the selection function. I will explain both approaches, beginning with the similarity semantics in Sect. 1.2, and moving on to the causal modeling semantics in Sect. 1.3.

1.2 Similarity semantics

Lewis (1973a, b) interprets points as possible worlds, and interprets the selection function in terms of comparative similarity or “closeness” among worlds. $f(\phi, w)$, Lewis suggests, is the set of “closest ϕ worlds to w ”—i.e., the set of worlds where ϕ is true which are at least as similar to w as any other world where ϕ is true.²

Lewis (1973b, pp. 58–60) argues that once we place the appropriate structural constraints on comparative similarity, the selection function must satisfy the following four constraints, and that furthermore, any selection function satisfying these constraints can be represented by an appropriately-structured similarity relation.

- (S1) If ϕ is true at w , then $f(\phi, w) = \{w\}$.
- (S2) ϕ is true at every point in $f(\phi, w)$.
- (S3) If ϕ is possible at w , then $f(\phi, w) \neq \emptyset$.
- (S4) If ψ entails ϕ and ψ is true at some points in $f(\phi, w)$, then $f(\psi, w) = \{v \in f(\phi, w) : \psi \text{ is true at } v\}$.

(S1) amounts to the requirement that every world be uniquely closest to itself, (S2) to the requirement that A be true at all the closest A worlds to w , (S3) to the requirement that if A is possible, some A worlds are closest to w , and (S4) to the requirement that the relation $\geq_w$ —where “ $w' \geq_w w''$ ” means that w' is closer to w than w'' is to w —has the structure of a total preordering.

While the structure of the similarity relation is uncontroversial, its content is a matter of considerable dispute. As Fine (1975) points out, it can’t look much like our ordinary notion of similarity. The sentence “if Nixon pressed the red button, a nuclear war would have started” is true. But a world in which Nixon presses the red

¹ Although this is the standard account of the relationship between selection and truth at a point, it is not the only possible account. We might say instead that $\phi \Box \rightarrow \psi$ is true at α if ψ is true at every point in $f(\phi, \alpha)$, false at α if ψ is false at every point in $f(\phi, \alpha)$, and indeterminate otherwise. We would then need a three-valued logic for assigning truth values to compounds of counterfactuals. Or we might say that $\phi \Box \rightarrow \psi$ is true at α if ϕ is true at *most* points in $f(\phi, \alpha)$, and false otherwise. For simplicity’s sake, I will ignore these alternatives and assume the standard account.

² To ensure that we can use the concept of closeness to define a selection function, we need the *Limit Assumption*, which forbids infinite sequences of closer and closer ϕ worlds to w with none closest. I will assume that the Limit Assumption holds; see Stalnaker (1981a, pp. 96–98) for a defense of it.

button and a nuclear war starts is much less similar to the actual world (in our ordinary sense of ‘similar’) than a world in which Nixon presses the red button, some mechanical malfunction prevents the rockets from launching, and life continues as usual. Subsequent literature is filled with sophisticated attempts to define an appropriate concept of similarity (Lewis 1979; Penczek 1997; Gundersen 2004; Kment 2006) and sophisticated counterexamples to them (Elga 2001; Edgington 2004; Hawthorne 2005; Krasner and Heller 1994; Wasserman 2006). I pass over these issues for the remainder of the article.

1.3 Causal models

The causal modeling account, defended by Galles and Pearl (1998), Pearl (2000), Hitchcock (2001), Woodward and Hitchcock (2003), Woodward (2003), and Hiddleston (2005), interprets points as causal models, and the selection function, in terms of a submodel relation. Different developments of the causal modeling framework differ in detail from one another, and from my own presentation, but they agree in basic outline.

To illustrate the concept of a causal model, it will be helpful to refer to an example from Pearl (2000, p. 207).

1.3.1 Execution

A firing squad (consisting of two executioners, X and Y) is about to shoot a prisoner. If the court orders the execution, this will cause the captain to signal to the executioners. Each executioner is posed to fire if, and only if, the captain signals. Either executioner’s shot by itself will be enough to kill the prisoner. As a matter of fact, the court does order the execution, so that the captain signals, the executioners shoot, and the prisoner dies.

In the course of explaining what a causal model is, I will construct a causal model EX to represent EXECUTION. In general, a causal model M is a triple $\langle \mathcal{V}, \mathcal{S}, \mathcal{A} \rangle$. I will explain each of these three elements below.

$\mathcal{V}$ is a finite set of *variables*, $\{V_1, V_2, \dots, V_n\}$. These variables pose questions about spatiotemporally circumscribed events. In EX , $\mathcal{V}$ contains five variables:

- U represents whether the court orders the execution.
- C represents whether the captain signals.
- X represents whether executioner X fires.
- Y represents whether executioner Y fires.
- D represents whether the prisoner dies.

Each $V \in \mathcal{V}$ comes with a range $\mathcal{R}(V)$ of possible values—that is, answers to the question posed by the variable. I assume that the answers to any given question are mutually exclusive and jointly exhaustive, and that no answer to one question entails an answer to any other (so that for instance, the correct answer to the question of whether the court ordered the execution does not logically entail the correct answer to the question of whether the captain signals). The simplest

variables have only two possible values, but I will allow for the possibility of variables that take on an arbitrary finite number of values. (A fully general causal modeling semantics would also allow for infinite-ranged variables, but I ignore these for the sake of technical simplicity.)

A sentence of the form $V_i = v_i$ means that the variable V_i takes on the value v_i . I will call such sentences *atomic*. In models like *EX*, where each variable is binary and corresponds to a ‘whether’ question, I will assume that the value 1 corresponds to the answer ‘yes’, while the value 0 corresponds to the answer ‘no’. I will use the name of the variable as shorthand for the sentence assigning value 1 to the variable, so that ‘ U ’ will sometimes stand for ‘ $U = 1$ ’, the atomic sentence stating that the court orders the execution. (I trust context to disambiguate situations where I am using a letter to represent a variable, and situations where I am using it to represent the corresponding atomic sentence.)

The second element of the triple, $\mathcal{S}$, is a set of *structural equations*, which specify relationships of direct causal dependence between variables. (For the purposes of this paper, I will assume that all such relationships are deterministic.) These structural equations can either be taken as primitive (Pearl 2000; Hiddleston 2005) or understood as representing certain privileged counterfactuals (Woodward 2003; Woodward and Hitchcock 2003; Hitchcock 2001); either way the basic idea is the same. For each variable V_i , $\mathcal{S}$ contains at most one equation of the form

$$V_i = f_i(PA_i).$$

PA_i is a subset of $\mathcal{V} - \{V_i\}$ —the set of V_i ’s *parents*. The equation $V_i = f_i(PA_i)$ says that V_i ’s value is causally determined by the values of the variables in PA_i , in the way stated by f_i .

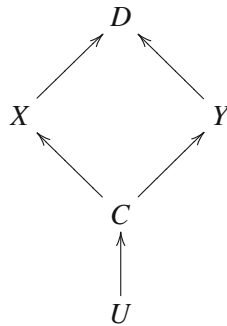
In *EX*, $\mathcal{S}$ contains the following structural equations.

$$\begin{aligned} C &= U \\ X &= C \\ Y &= C \\ D &= \max\{X, Y\} \end{aligned}$$

The first of these equations has C on the lefthand side, and U on the right. This means that C —whether the captain signals—depends causally on U —whether the court orders the execution—and not on any of the other variables in the model. The equation gives the further information that C takes on the same value as U —that is, the captain will signal if the court orders the execution, and won’t signal if the court doesn’t order the execution. Likewise, the middle two equations say that each of the executioners will fire if and only if the captain signals, while the final equation says that the prisoner will die if and only if at least one of the executioners fires.

Notice that there is no structural equation for U . This is because the causes of U are not represented anywhere in the model; in technical terms, U is *exogenous*. The remaining variables in *EX* are *endogenous*: their values are causally determined by the values of other variables in the model. A variable is exogenous if there is a structural equation corresponding to it, and endogenous otherwise.

The variables and structural equations for a causal model can be represented by a graph. The graph for *EX* is as follows.



Each node of the graph represents a variable, while an arrow pointing from variable V to variable V' represents V as a parent of V' . The graph is useful for visualising a number of other relations defined in terms of the parent relation. Naturally, the converse of the parent relation is the *child* relation: if there is an arrow pointing from V to V' , then V' is the child of V . The ancestral of the parent relation is the *ancestor* relation: if one can get from V to V' by following a series of arrows from tail to tip, then V is an ancestor of V' . (Thus, for instance, Y 's ancestors are C and U .) The ancestral of the child relation (i.e., the converse of the ancestor relation) is the *descendant* relation: if one can get from V to V' by following a series of arrows tip to tail, then V is a descendant of V' . Say that V' is a *nondescendant* of V iff $V' \neq V$ and V' is not a descendant of V . (In my terminology, a variable may be neither a descendant nor a nondescendant of itself. I will restrict my attention to *recursive* causal models, in which no variable is its own descendant. In graphical terms, a recursive model is one in which one can never leave a variable and then return to the same variable by following a series of arrows from tail to tip.

The third member of the triple in a causal model, $\mathcal{A}$, is a function assigning a value $\mathcal{A}(V_i)$ to every $V_i \in \mathcal{V}$ in a way consistent with the structural equations. In *EX*, $\mathcal{A}$ assigns value 1 to each of the variables.

My characterization of $\mathcal{A}$ incorporates a few choices which are worth highlighting here. First, I count $\mathcal{A}$ as a part of the model, whereas Galles and Pearl (1998) and Pearl (2000) individuate models by $\mathcal{V}$ and $\mathcal{S}$ alone. I have made this choice to keep the analogy between models and worlds as close as possible: my models represent not just dependencies between possible events, but also facts about which events occurred. Second, I have made $\mathcal{A}$ an assignment of values to all the variables, rather than just the exogenous ones. In indeterministic and nonrecursive models, the values of exogenous variables can be recovered from the values of all variables, but the converse is not always true. I have chosen the assignment function so that it carries as much information as possible. It will be easier to ignore the extra information, should it prove unnecessary in later inquiries, than to recover it, should it prove necessary.

We can use causal models to partially define a selection function. I will illustrate this concept by showing how to evaluate the following counterfactual in *EX*: “If court had not ordered the execution, but executioner X had shot anyway, the prisoner would have died”—or formalized, $(\neg U \wedge X) \square \rightarrow D$.

The first step is to define an *intervention*. In general, where A is a sentence of the form $V_1 = c_1 \wedge \dots \wedge V_m = c_m$, the intervention on M that sets each $V_i \in \mathcal{V}_A = \{V_1, \dots, V_m\}$ equal to c_i

replaces the structural equation $V_i = f_i(pa_i)$ with the structural equation $V_i = c_i$, if V_i is endogenous, and

sets the value of V_i to c_i , if V_i is exogenous.

In the example of the counterfactual $(\neg U \wedge X) \square \rightarrow D$, we want to consider an intervention that sets $U = 0$ and $X = 1$.

Intuitively, an intervention takes each $V_i \in \mathcal{V}_A$, removes the mechanism that determines V_i 's value, and replaces the old mechanism with a new one, while leaving the other causal mechanisms in M and the values of V_i 's nondescendants intact.³ In EXECUTION, one might intervene to set $U = 0$ and $X = 1$ by tampering with the court's memos so that no execution is ordered, and bribing executioner X to shoot regardless of whether the captain signals.

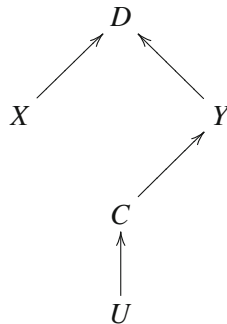
An intervention on a model $M = \langle \mathcal{V}, \mathcal{S}, \mathcal{A} \rangle$ generates a *submodel* $M_A = \langle \mathcal{V}, \mathcal{S}_A, \mathcal{A}_A \rangle$ —the model that would result from performing the appropriate intervention on M . (Notice that the $\mathcal{V}$ in M_A is exactly the same as the $\mathcal{V}$ in M ; interventions have no effect on which variables are part of a model.)

To generate $EX_{\neg U \wedge X}$ —the submodel of EX that results from setting $U = 0$ and $X = 1$ —we begin by modifying $\mathcal{S}$ to get $\mathcal{S}_{\neg U \wedge X}$.

$$\begin{aligned} C &= U \\ X &= 1 \\ Y &= C \\ D &= \max(X, Y) \end{aligned}$$

I have simply replaced the equation $X = C$ with the equation $X = 1$, and left all of the other equations in $\mathcal{S}$ unchanged.

$\mathcal{V}$ and $\mathcal{S}_{\neg U \wedge X}$ are depicted in the graph below. (Notice how I have removed the arrow between C and X to show that the causal connection between C and X is severed.)



The only remaining step is to define our new assignment function $\mathcal{A}_{\neg U \wedge X}$. Where M_A is a submodel of M , and $V_i \in \mathcal{V}$,

³ See Woodward (2003, pp. 98–114) for a detailed discussion of the concept of an intervention.

If A contains the conjunct $V_i = c_i$, then $\mathcal{A}_A(V_i) = c_i$.

If V_i is a nondescendant of every variable in $\mathcal{V}_A$, then $\mathcal{A}_A(V_i) = \mathcal{A}(V_i)$.

Otherwise, $\mathcal{A}_A(V_i)$ is determined by the structural equations in $\mathcal{S}_A$, together with the values of other variables.

The basic idea is to change the values of the variables in $\mathcal{V}_A$ by intervention, leave the variables causally upstream from them alone, and let the effects of the intervention propagate causally downstream according to the structural equations. In the case of *EX*,

$$\mathcal{A}_{\neg U \wedge X}(U) = 0$$

$$\mathcal{A}_{\neg U \wedge X}(C) = 0$$

$$\mathcal{A}_{\neg U \wedge X}(X) = 1$$

$$\mathcal{A}_{\neg U \wedge X}(Y) = 0$$

$$\mathcal{A}_{\neg U \wedge X}(D) = 1$$

In general, $f(A, M)$ is the set of submodels of M generated by interventions that bear the appropriate relationship to A . (So far, we have considered only antecedents for which $f(A, M)$ is a singleton, but that will soon change.) In *EXECUTION*, $f(EX, \neg U \wedge X)$ contains only one model, and in this model, D is true. So the entire conditional $(\neg U \wedge X) \Box \rightarrow D$ is true in *EX*.

2 Comparisons

The similarity account and the causal modeling account give different ways of cashing out the concept of a selection function. What are the consequences of adopting one rather than the other? I will compare and contrast what the two views have to say about the truth conditions of counterfactuals, and what they have to say about the validity of inferences containing counterfactuals.

2.1 Truth conditions

Read as complete accounts of the truth conditions of counterfactuals, the causal modeling and similarity approaches conflict.⁴ The similarity approach holds that the truth values of counterfactuals depend on “all-over similarity” among possible worlds, and that assessing similarity requires “taking account of all the respects of similarity and differences and balancing them off against one another” (Lewis 1973a, p. 559). The causal modeling account, on the other hand, seems to assign

⁴ One can make pragmatic use of either approach without reading it as a complete account of the truth conditions of counterfactuals. For instance, one might think the similarity account is heuristically useful, but mistaken whenever it conflicts with the causal modeling account. Or one might accept (Hall 2007, p. 110)’s claim that the causal modeling framework is “a useful means for selectively representing aspects of an antecedently understood counterfactual structure”, but does not provide truth conditions for all counterfactuals of interest. In this article, however, I take both accounts at face value.

truth values to counterfactuals based only on the local features of circumscribed parts of the world. Nonetheless, one could in principle adopt a Lewis-style similarity account in which the points were not worlds, but circumscribed parts of the world (e.g., situations), or a structural equations model of the entire world.

The two accounts of counterfactuals are also loosely connected to different views about Humean Supervenience—the thesis that the fundamental features of the world are non-nomological, and that counterfactual dependence, causation, laws of nature, and chances all supervene on these fundamental, non-nomological features.⁵ Lewis (1986a, b, 1994), the chief proponent of the similarity account, espouses Humean Supervenience, while proponents of the causal modeling account often voice anti-reductionist sentiments (Pearl, 2000, pp. 21–22, pp. 134–135) (Woodward 2003, pp. 104–107). It is easy to see why a commitment to Humean Supervenience might motivate a person to choose the similarity account over the causal modeling account: similarity is not a nomologically loaded concept, but structural equations are difficult to understand unless one attributes some kind of causal or nomological significance to them. Nonetheless, some authors (Schaffer 2004; Edgington 2004) defend anti-Humean versions of the similarity account, and in principle, one could adopt a reductive, non-nomological interpretation of the structural equations.

The most striking difference between the similarity and causal modeling accounts is that the similarity account does, and existing causal modeling accounts do not, allow us to assign truth values to arbitrary counterfactuals. Galles and Pearl consider a language in which the only expressible counterfactuals have the form $(A_1 \wedge \dots \wedge A_n) \Box \rightarrow (C_1 \wedge \dots \wedge C_n)$, where $A_1 \dots A_n$ and $C_1 \dots C_n$ are atomic. Halpern and Hiddleston provide semantics for a slightly more extended language, in which the only expressible counterfactuals have the form $(A_1 \wedge \dots \wedge A_n) \Box \rightarrow C$, where $A_1 \dots A_n$ are atomic, and C is an arbitrary Boolean combination of atomic sentences.⁶

The causal modeling account has a second, related limitation: it only partially defines a selection function. $f(A, M)$ is well defined only when A is a conjunction of atomic sentences. In Sect. 3, I will extend the causal modeling account to provide a more extensive definition of the selection function, and truth conditions for a wider range of counterfactuals. But first, I will compare the logical features of the similarity and causal modeling accounts as I have cashed them out so far.

⁵ Lewis (1986b, p. 111) glosses Humean Supervenience this way: “if two worlds match perfectly in all matters of particular fact, they match perfectly in all other ways too—in modal properties, laws, causal connections, chances.”

⁶ Some versions of the similarity account also limit the range of permissible antecedents. One way of defining the selection function, proposed by Lewis (1979), involves mapping each antecedent A onto a stretch of time t_A , where A “is entirely about” events that happen in t_A . Another type of proposal, defended by Schaffer (2004) and Edgington (2004), defines the selection function in terms of causal independence. Not all propositions are about particular stretches of time, and not all propositions describe events that are eligible to stand in relations of causal dependence or independence. Where A_1 and A_2 are each about particular stretches of time, $A_1 \vee A_2$ need not be. Likewise, where A_1 and A_2 each describe events that stand in relations of causal dependence and independence, $A_1 \vee A_2$ need not.

2.2 Validity

Similarity accounts of counterfactual dependence are characterized by the “system-of-spheres” logics set out in (Lewis 1973b) and (Lewis 1981). These logics share the following core set of rules and axioms, collectively known as V.⁷

2.2.1 Rules

- (R1) If $\vdash \phi$ and $\vdash \phi \supset \psi$, then $\vdash \psi$
 (R2) If $\vdash ((\psi_1 \wedge \dots \wedge \psi_n) \supset \theta)$,
 then $\vdash ((\phi \Box \rightarrow \psi_1) \wedge \dots \wedge (\phi \Box \rightarrow \psi_n)) \supset (\phi \Box \rightarrow \theta)$

2.2.2 Axioms

- (V1) All truth-functional tautologies.
 (V2) $\phi \Box \rightarrow \phi$
 (V3) $((\phi \Box \rightarrow \psi) \wedge (\psi \Box \rightarrow \phi)) \supset ((\phi \Box \rightarrow \theta) \equiv (\psi \Box \rightarrow \theta))$
 (V4) $((\phi \vee \psi) \Box \rightarrow \phi) \vee ((\phi \vee \psi) \Box \rightarrow \psi) \vee (((\phi \vee \psi) \Box \rightarrow \theta) \equiv ((\phi \Box \rightarrow \theta) \wedge (\psi \Box \rightarrow \theta)))$
 (V5) $(\phi \Box \rightarrow \psi) \supset (\phi \supset \psi)$

The class of counterfactual logics satisfying selection conditions (S1)-(S4) is characterized by the system VC, which consists of V plus the additional axiom

$$(C) (\phi \wedge \psi) \supset (\phi \Box \rightarrow \psi).$$

Galles and Pearl (1998) argue that VC is sound and complete over the class of recursive causal models. But their completeness proof applies only to the fragment of the language in which all counterfactuals take the form $(A_1 \wedge \dots \wedge A_n) \Box \rightarrow (C_1 \wedge \dots \wedge C_n)$, where $A_1 \dots A_n$ and $C_1 \dots C_n$ are atomic. The proof also assumes that the variables have a built-in causal ordering, and that the language contains axioms specifying this ordering.

These assumptions can and should be relaxed. In Sect. 3, I will develop a semantics for a language of counterfactuals $\mathcal{L}$ that permits both arbitrary Boolean compounds of atomic sentences in the antecedents of counterfactuals, and counterfactuals in the consequents of counterfactuals. The variables in $\mathcal{L}$ will have no built-in causal ordering. In Sect. 4, I will provide a sound and complete axiomatization for $\mathcal{L}$.

3 Expanding the language

We can and should expand the causal modeling language to allow for a wider range of antecedents and consequents. Expanding the range of allowable

⁷ I use the axiomatization in (Lewis 1981), since it is the one that has been taken up in the causal modeling literature.

consequents is simple, and requires no extra semantic apparatus—only more sophisticated sentence structure. When we consider, at α , what would have happened in counterfactual circumstances A (for instance, when we ask of EX , “what if the court had not ordered the execution?”), we need not confine our answers to conjunctions of atomic sentences (“the prisoner would have survived”). We can also consider complex sentences as answers (“either the prisoner would have survived or executioner X would have shot anyway”; “if executioner X had shot anyway, then the prisoner would have died”). The selection semantics already provides enough equipment to evaluate counterfactuals with complex consequents, since sentences of all types have truth values at points in $f(A, \alpha)$.

We should also extend the definition of the selection function to expand the range of allowable antecedents. This will allow the causal modeling account to answer questions what would happen in logically complex counterfactual circumstances. There are good reasons to ask such questions. Sartorio (2006) argues persuasively that disjunctions can be causes, and if she is right, we should expect propositions to depend counterfactually on disjunctions. Furthermore, if, as Fodor (1974) conjectures, mental states are multiply realizable by neural states, then being in pain will consist not in being in any particular neural state N , but in being in either some neural state N_1 , or some distinct neural state in N_2 or And suppose that we want to know what would happen if a particular subject were in pain. We should be able to answer this question by determining what would happen if the subject were in N_1 or N_2 or More complicated reductions of higher-level states are possible: perhaps being in pain is being both in state N_1 and in either state N_2 or state N_3 . In order to understand disjunctive causes, or to connect counterfactuals about lower-level events to counterfactuals about higher-level events, we need a way of understanding counterfactuals with complex antecedents. (Although it would be worthwhile to provide semantics for counterfactuals with counterfactual antecedents, in this essay I will concentrate only on counterfactuals with Boolean antecedents.)

I will therefore define a semantics for an extended causal modeling language, $\mathcal{L}$.

Vocabulary $\mathcal{L}$ contains

two propositional constants, $\top$ and $\perp$.

An arbitrary finite number of variables with finite ranges of arbitrary size. Each variable V is syntactically marked with a range $\mathcal{R}(V)$.

connectives $\wedge, \vee, \neg, \supset$, and $\Box \rightarrow$

left and right parentheses

Well-formed formulae A wff of $\mathcal{L}$ is defined as follows.

$V = v \in \mathcal{L}$ when $v \in \mathcal{R}(V)$.

If $\phi \in \mathcal{L}$ and $\psi \in \mathcal{L}$, then $\neg\phi \in \mathcal{L}$, $(\phi \wedge \psi) \in \mathcal{L}$, $(\phi \vee \psi) \in \mathcal{L}$, and $(\phi \supset \psi) \in \mathcal{L}$.

If $A \in \mathcal{L}$, $\phi \in \mathcal{L}$, and A contains no instances of $\Box \rightarrow$, then $(A \Box \rightarrow \phi) \in \mathcal{L}$.

Valuation function Let each model M be assigned a *valuation function* v_M , which assigns 1 to all propositions that are true in M and 0 to all propositions that are false in M .

$$v_M(V = v) = \begin{cases} 1 & \text{if } \mathcal{A}(V) = v \\ 0 & \text{else} \end{cases}$$

$$v_M(\neg\phi) = 1 - v_M(\phi)$$

$$v_M(\phi \wedge \psi) = v_M(\phi)v_M(\psi)$$

$$v_M(\phi \vee \psi) = \max\{v_M(\phi), v_M(\psi)\}$$

$$v_M(\phi \Box \rightarrow \psi) = \begin{cases} 1 & \text{if } v_M(\psi) = 1 \text{ for every } M' \in f(\phi, M) \\ 0 & \text{else} \end{cases}$$

I will use the Greek letters ϕ , ψ , and θ for arbitrary formulae, the italicized Roman letters A , B , and C for counterfactual-free formulae, and the boldface Roman letters **A**, **B** and **C** for conjunctions of atomic formulae in which no two conjuncts mention the same variable.

3.1 Logically complex consequents

Let us begin, then, by considering counterfactuals with logically complex consequents. Much of useful work has already been done by Halpern (2000), who extends the axiomatization of Galles and Pearl (1998) to counterfactuals with arbitrary Boolean consequents. But what about counterfactuals with counterfactual consequents?

The semantics for counterfactuals with counterfactual consequents is already implicit in the selection approach. $\phi \Box \rightarrow (\psi \Box \rightarrow \theta)$ is true in M if $\psi \Box \rightarrow \theta$ is true in every $M' \in f(\phi, M)$ and false otherwise. But this simple corollary of the selection account has a surprising consequence: it makes *modus ponens* invalid, so that ϕ and $\phi \Box \rightarrow \psi$ do not jointly entail ψ .

To see why *modus ponens* fails, consider the following counterfactual, as evaluated in the model EX .⁸

If executioner X had fired, then (even) if the captain had not signalled, the prisoner would have died.

$$X \Box \rightarrow (\neg C \Box \rightarrow D)$$

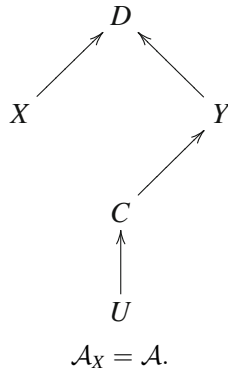
On the causal modeling semantics for nested counterfactuals, this counterfactual comes out true. To see this, begin by considering the submodel $EX_X = \langle V, \mathcal{S}_X, \mathcal{A}_X \rangle$

⁸ Astute readers will notice similarities between my counterexample to *modus ponens* and the counterexamples famously proposed by McGee (1985). McGee formulates counterexamples for both an indicative interpretation of the conditional—in which points are probability functions, truth at a point for nonconditional sentences is equated with high probability, and $f(A, P)$ is the probability function that results from conditionalizing P on A —and a counterfactual interpretation of the conditional—where the semantics agrees with Lewis's in all relevant respects except centering. On McGee's two interpretations of the conditional, and on mine, whenever A is true at a point α , $f(A, \alpha)$ must contain a point β that assigns the same truth values to all nonconditional sentences as α . But the features of β that ground the selection function are slightly different from the features of α that ground the selection function; thus α and β assign different truth values to conditionals. Although *modus ponens* is valid for conditionals with nonconditional consequents, it is not valid for conditionals in general.

generated by intervening on EX to set $X = 1$. $\mathcal{S}_X$ contains the following structural equations:

$$\begin{aligned} C &= U \\ X &= 1 \\ Y &= C \\ D &= \max\{X, Y\} \end{aligned}$$

These structural equations, and the variables in them, are depicted in the following graph:

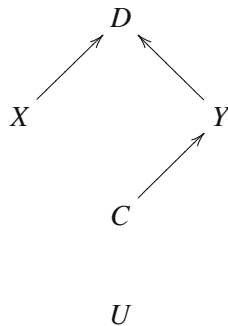


Our original counterfactual $X \square \rightarrow (\neg C \square \rightarrow D)$ will be true in EX if and only if its consequent $\neg C \square \rightarrow D$ is true in EX_X . To check the truth value of $\neg C \square \rightarrow D$ in EX_X , we must consider the model $(EX_X)_{\neg C} = \langle V, (\mathcal{S}_X)_{\neg C}, (\mathcal{A}_X)_{\neg C} \rangle$, which results from intervening on EX_X to set $C = 0$. If D is true in this model, then $\neg C \square \rightarrow D$ is true in EX_X , and so the original counterfactual $X \square \rightarrow (\neg C \square \rightarrow D)$ is true in EX .

$(\mathcal{S}_X)_{\neg C}$ contains the following structural equations:

$$\begin{aligned} C &= 0 \\ X &= 1 \\ Y &= C \\ D &= \max\{X, Y\} \end{aligned}$$

These structural equations, and the variables in them, are depicted in the following graph:

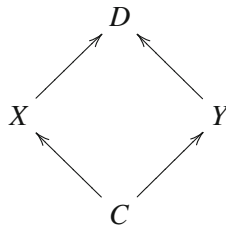


It is simple to verify that $(\mathcal{A}_X)_{\neg C}(D) = 1$. Therefore, $X \Box \rightarrow (\neg C \Box \rightarrow D)$ is true in EX .

Plainly, X is true in EX . But we can verify that $(\neg C \Box \rightarrow D)$ is false in EX . Consider the model $EX_{\neg C} = \langle V, \mathcal{S}_{\neg C}, \mathcal{A}_{\neg C} \rangle$, which results from intervening on EX to set $C = 0$. $\mathcal{S}_{\neg C}$ contains the following structural equations:

$$\begin{aligned} C &= 0 \\ X &= C \\ Y &= C \\ D &= \max\{X, Y\} \end{aligned}$$

These structural equations, and the variables in them, are depicted in the following graph:



U

And the structural equations ensure that $\mathcal{A}_{\neg C}(D) = 0$. So $\neg C \Box \rightarrow D$ is false in EX .

Putting this all together, we have the result that $X \Box \rightarrow (\neg C \Box \rightarrow D)$ and X are both true in EX , while $\neg C \Box \rightarrow D$ is false in EX . So *modus ponens* is invalid.

This example shows that the causal modeling semantics violates *weak centering*—the assumption that if ϕ is true at α , then $\alpha \in f(\phi, \alpha)$. (In Lewis's similarity semantics, centering is a consequence of condition (S1).) Intervening to set $V = v$ may change the model, even if $V = v$ was true to begin with. The change is subtle enough not to show up in counterfactuals with true antecedents and non-counterfactual consequents, since intervening to set a variable to its actual value won't change the values of any variables. But the change shows up in counterfactuals with true antecedents and counterfactual consequents, since intervening to set a variable to its actual value will change the counterfactual relationships among variables.

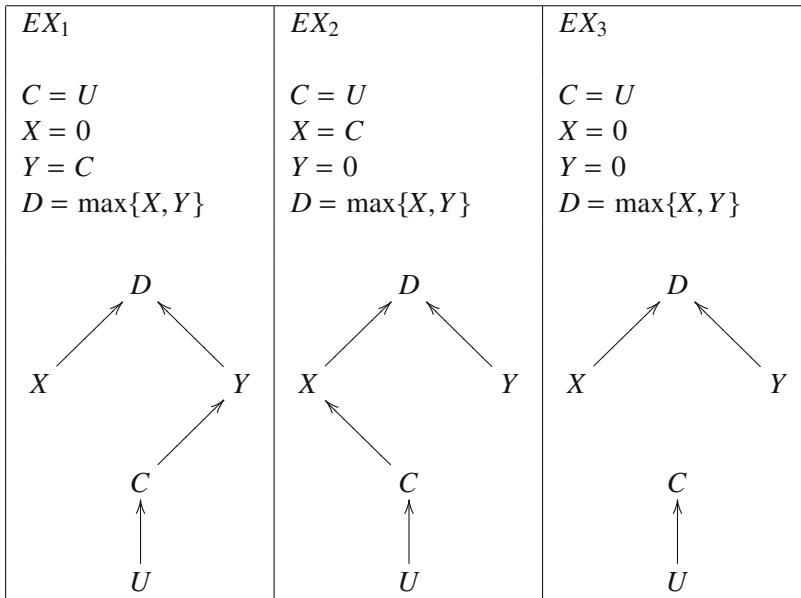
3.2 Logically complex antecedents

It is time to move on to counterfactuals with logically complex antecedents. Let us begin with a simple example: What is the truth value of the following counterfactual in EXECUTION?

If executioner X hadn't fired or executioner Y hadn't fired, then the prisoner would not have died.

$$(\neg X \vee \neg Y) \Box \rightarrow \neg D$$

The antecedent of this counterfactual fails to correspond to a unique intervention. There are three different and equally good interventions that make it true: one that sets $X = 0$, another that sets the $Y = 0$, and a third that sets both $X = 0$ and $Y = 0$. Thus, $f(\neg X \vee \neg Y, EX)$ contains three equally good candidate submodels.



It is easy to see that in EX_1 and EX_2 , $D = 1$. Thus, $(\neg X \vee \neg Y) \Box \rightarrow \neg D$ is false in EX .

Here a potential objection arises: on what grounds do I claim that all three submodels are equally good? Perhaps if either X or Y hadn't fired, X would not have fired, and Y would have fired anyway. In that case, EX_1 is uniquely the best submodel, and I have mischaracterized $f(\neg X \vee \neg Y, EX)$.

This objection would be effective if EX were merely a useful way of encoding counterfactual information which is ultimately grounded in other sources. But I propose to take causal modeling accounts seriously, as complete stories about the semantics of counterfactuals. And EX seems to give us no grounds for choosing among the three submodels EX_1 , EX_2 , and EX_3 . In the absence of any good argument to the contrary from defenders of causal modeling, I will assume that all of them are on a par, and therefore all of them belong to $f(\neg X \vee \neg Y, EX)$.

Note that this is a choice point. Hüber (manuscript) adopts a different response to this choice point: he suggests a way of breaking apparent ties. For Hüber, the further upstream an intervention on an endogenous variable occurs, the more it makes for dissimilarity between the original model and the submodel. Hüber also stipulates that agreement about the values of exogenous variables makes for similarity among models. He proves that if these constraints are cashed out properly, then the logic V is sound and complete over the causal modeling semantics. Hüber's account has a cost: "the limited set of counterfactuals we get from the structural equations is not

enough to represent the one relevant modality of counterfactuality” (Hüber, manuscript, 17). Some logicians may hold that the ability to hold on to Lewis’s logic V is worth the price; however, I will explore what happens when we take the opposite direction at this choice point.

A fully general account of counterfactuals with complex antecedents requires a procedure for generating $f(A, M)$ based on a model M and an arbitrary Boolean compound of atomic sentences A . Luckily, a proposal by Fine (forthcoming) can help. Following Fine, we can cash out the selection function by appealing the concept of a *state space*: a possible world is composed of states, which serve as truthmakers for sentences. A causal model is a type of state space.

Each atomic sentence is made true by a state consisting of a variable plus its value; for instance, in EX , the sentence “ X ” or “ $X = 1$ ” is made true by the state of X ’s firing (the variable X taken together with the value 1). Furthermore, each conjunction of atomic sentences is made true by a state which is a fusion of simpler states; for instance, in EX , the sentence “ $X \wedge \neg Y$ ” corresponds to the state that consists of X ’s firing and Y ’s not firing. (Call the fusion of states s and t $s \sqcup t$.) In formal terms, the following principles govern states and their fusions.

Atomic States For each variable V and value v in V ’s range, there is a state $s[V = v]$.

Fusion Where $V_1, \dots, V_n$ are distinct variables, there is a state $s[V_1 = v_1] \sqcup \dots \sqcup s[V_n = v_n]$.

Exclusion Where $v \neq v'$, $s[V = v]$ and $s[V = v']$ have no fusion.

Null State There is a null state $\top$, such that for every state s , $s \sqcup \top = s$

Identity For every state s , $s \sqcup s = s$.

Transitivity If $s \sqsubseteq t$ and $t \sqsubseteq s'$, then $s \sqsubseteq s'$.

Fusion-Containment If $s \sqsubseteq t$ and $s' \sqsubseteq t$, then $s \sqcup s' \sqsubseteq t$.

Extensionality If for every state s , $s \sqsubseteq t$ iff $s \sqsubseteq t'$, then $t = t'$.

A couple of notes are in order. First, I count a state as existing if it could be instantiated, regardless of whether it actually is instantiated. Thus, even though X fires in EX , there exists in EX a state of X ’s not firing. Second, there are no disjunctive states that correspond directly to sentences like “ $X \vee \neg Y$ ”, and likewise, no negative states that correspond directly to sentences like “ $\neg(X \wedge Y)$ ”.

In light of the second point, how should we define truthmaking for disjunctions, negations, and other complex sentences? We can borrow a set of rules from Fine (forthcoming, p. 10). (I will write “ $s \models \phi$ ” for “ s makes ϕ true.”).⁹

$$(V^+) s \models V = v \text{ iff } s = s[V = v]$$

$$(V^-) s \models V \neq v \text{ iff } s = s[V = v'] \text{ for some } v' \neq v.$$

⁹ I’ve made few modifications to Fine’s rules. I appeal only to verification conditions, whereas Fine appeals to both verification conditions and falsification conditions. And I’ve added verification conditions for atomic sentences and the material conditional.

- $(\neg) s \models \neg \neg \phi$ iff $s \models \phi$
 $(\top) s \models \top$ iff $s = \top$
 $(\perp) s \models \perp$ never
 $(\vee^+) s \models \phi \vee \psi$ iff $s \models \phi$ or $s \models \psi$ or $s = t \sqcup r$ where $t \models \phi$ and $r \models \psi$
 $(\vee^-) s \models \neg(\phi \vee \psi)$ iff $s = t \sqcup r$ where $t \models \neg \phi$ and $r \models \neg \psi$
 $(\wedge^+) s \models \phi \wedge \psi$ iff $s = t \sqcup r$ where $t \models \phi$ and $r \models \psi$
 $(\wedge^-) s \models \neg(\phi \wedge \psi)$ iff $s \models \neg \phi$ or $s \models \neg \psi$ or $s = t \sqcup r$ where $t \models \neg \phi$ and $r \models \neg \psi$
 $(\supset^+) s \models \phi \supset \psi$ iff $s \models \neg \phi$ or $s \models \psi$ or $s = t \sqcup r$ where $t \models \neg \phi$ and $r \models \psi$
 $(\supset^-) s \models \neg(\phi \supset \psi)$ iff $s = t \sqcup r$ where $t \models \phi$ and $r \models \neg \psi$

We are now ready to define the selection function. Each state corresponds to an intervention: $\top$ corresponds to a null intervention, which makes no changes to the original model, and a state $s[V_1 = v_1] \sqcup \dots \sqcup s[V_n = v_n]$ corresponds to the intervention of setting $V_1 = v_1$ and $\dots V_n = v_n$. We can say that $f(\phi, M)$ is the set of submodels generated by interventions on M corresponding to states that verify ϕ .

This definition of the selection function has one important upshot, which is worth highlighting. The concept of truthmaking generates a sense of equivalence which is narrower than the ordinary classical one, and which I will call *exact* equivalence, following Fine. Two sentences are exact equivalents iff they have the same truthmakers. Not every pair of classically equivalent sentences is exactly equivalent; for instance, $(\neg X \wedge Y) \vee (\neg X \wedge \neg Y)$ is classically, but not exactly, equivalent to $\neg X$. Although exact equivalents can be freely substituted in the antecedents of counterfactuals, classical equivalents cannot. For instance, in EX , it is true that $\neg X \Box \rightarrow D$ (since D is true in the one $\neg X$ submodel of EX), but false that $((\neg X \wedge Y) \vee (\neg X \wedge \neg Y)) \Box \rightarrow D$ (since D is false in the $\neg X \wedge Y$ submodel of EX).

4 Logic

The following system of rules and axioms (call it CM, for ‘Causal Modeling’) is sound and complete over the semantics for $\mathcal{L}$. (Recall that Greek letters ϕ, ψ , and θ stand for arbitrary formulae, italicized Roman letters A, B , and C stand for counterfactual-free formulae, and the boldface Roman letters **A**, **B** and **C** stand for conjunctions of atomic formulae in which no two conjuncts mention the same variable.)¹⁰

4.1 Rules

- (R1)** If $\vdash \phi$ and $\vdash \phi \supset \psi$, then $\vdash \psi$
(R2) If $\vdash ((\psi_1 \wedge \dots \wedge \psi_n) \supset \theta)$,
 then $\vdash ((\phi \Box \rightarrow \psi_1) \wedge \dots \wedge (\phi \Box \rightarrow \psi_n)) \supset (\phi \Box \rightarrow \theta)$

¹⁰ I owe much to the axiomatization given by Halpern (2000). The first eight axioms are similar in spirit to Halpern’s axioms; the last three concern counterfactuals with complex antecedents and consequents, and are entirely my own.

4.2 Axioms

- (CM1) All truth-functional tautologies.
 (CM2) $\bigvee_{v_i \in \mathcal{R}(V_i)} V_i = v_i$
 (CM3) $\neg(V = v \wedge V = v')$, where $v \neq v'$
 (CM4) $(\mathbf{A} \square \rightarrow \phi) \vee (\mathbf{A} \square \rightarrow \neg \phi)$
 (CM5) $\neg(\mathbf{A} \square \rightarrow \perp)$
 (CM6) $((\mathbf{A} \wedge V = v) \square \rightarrow B) \supset (\mathbf{A} \square \rightarrow (V = v \supset B))$, where none of $\mathbf{A}$'s conjuncts mentions V .
 (CM7) $\mathbf{A} \square \rightarrow \mathbf{A}$
 (CM8) $(d_1(V_1, V_2) \wedge \dots \wedge d_{n-1}(V_{n-1}, V_n)) \supset \neg d_n(V_n, V_1)$ where each $d_1(V_i, V_{i+1})$ is a conjunction of the form $((\mathbf{A}_i \wedge V_i = v_i) \square \rightarrow V_{i+1} = v_{i+1}) \wedge ((\mathbf{A}_i \wedge V_i = v'_i) \square \rightarrow V_{i+1} \neq v_{i+1})$ and none of $\mathbf{A}_i$'s conjuncts mentions V_i .
 (CM9) $((A \vee B) \square \rightarrow \phi) \equiv ((A \square \rightarrow \phi) \wedge (B \square \rightarrow \phi) \wedge ((A \wedge B) \square \rightarrow \phi))$
 (CM10) $(A \square \rightarrow \phi) \equiv (A' \square \rightarrow \phi)$ where A and A' are exact equivalents.
 (CM11) $((\mathbf{A} \wedge \mathbf{B} \wedge \mathbf{C}) \square \rightarrow D) \equiv ((\mathbf{A} \wedge \mathbf{B}') \square \rightarrow ((\mathbf{B} \wedge \mathbf{C}) \square \rightarrow D))$ where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ share none of their variables, and $\mathbf{B}$ and $\mathbf{B}'$ share all their variables.

Rules (R1) and (R2), as well as axioms (CM1) and (CM7), should already be familiar from Lewis's logic $\mathbf{V}$. Axioms (CM2) and (CM3) regulate the variables: they ensure that each variable takes one, and only one, value. (CM4) and (CM5) are versions of Conditional Excluded Middle and Conditional Non-Contradiction, restricted to counterfactuals with a particular type of antecedent (roughly, the type of antecedent that is made true by exactly one state). (CM6) is a version of the axiom that Galles and Pearl (1998) and Halpern (2000) call "Composition"; Galles and Pearl (1998, p. 162) gloss it as claiming "that if we force a variable ... to a value that it would have had without our intervention, then our intervention will have no effect on the other variables in the system". It allows us to derive *modus ponens* for counterfactuals with non-counterfactual consequents. (CM8) ensures that all models for the system will be recursive by forbidding cyclical patterns of counterfactual dependence. (CM9) states that a counterfactual with a disjunctive antecedent $A \vee B$ is equivalent to a conjunction of simpler counterfactuals with antecedents A , B , and $A \wedge B$.¹¹ (CM10) says that exact equivalents can be substituted freely in the antecedents of counterfactuals. Finally, (CM11) says that certain right-nested conditionals (roughly, those where each antecedent is made true by exactly one state) are equivalent to simple conditionals. Note that (CM11) entails a special case of the so-called "Import-Export" principle: where $\mathbf{A}$ and $\mathbf{B}$ share none of their variables, $(\mathbf{A} \wedge \mathbf{B}) \square \rightarrow C$ is equivalent to $\mathbf{A} \square \rightarrow (\mathbf{B} \square \rightarrow C)$.

CM is not an extension of $\mathbf{V}$, for it invalidates rules (V3), (V4), and (V5). One can verify that the following instances of (V3), (V4), and (V5) are false in the model *EX*.

¹¹ (CM9) ensures that CM validates the controversial rule *Simplification of Disjunctive Antecedents* (SDA) $(\phi \vee \psi) \square \rightarrow \theta \vdash (\phi \square \rightarrow \theta) \wedge (\psi \square \rightarrow \theta)$. Although Nute (1978) defends SDA, most authors reject the principle on the grounds that it leads to undesirable consequences when taken together with the substitution of classical equivalents (see Fine 1975; see Loewer 1976). But the concept of exact equivalence gives us independent reason to reject the substitution of classical equivalents.

- (V3) $((\neg X \Box \rightarrow ((\neg X \wedge Y) \vee (\neg X \wedge \neg Y))) \wedge (((\neg X \wedge Y) \vee (\neg X \wedge \neg Y)) \Box \rightarrow \neg X))$
 $\supset ((\neg X \Box \rightarrow D) \equiv ((\neg X \wedge Y) \vee (\neg X \wedge \neg Y)) \Box \rightarrow D)$
- (V4) $((\neg X \vee \neg Y) \Box \rightarrow \neg X) \vee ((\neg X \vee \neg Y) \Box \rightarrow \neg Y) \vee (((\neg X \vee \neg Y) \Box \rightarrow D) \equiv$
 $((\neg X \Box \rightarrow D) \wedge (\neg Y \Box \rightarrow D)))$
- (V5) $(X \Box \rightarrow (\neg C \Box \rightarrow D)) \supset (X \supset (\neg C \Box \rightarrow D))$

The difference between CM and systems-of-spheres logics is also noticeable at the level of the selection function. The selection function for CM violates (S1), since $EX \notin f(X, EX)$. It also violates (S4), since $(\neg X \wedge Y) \vee (\neg X \wedge \neg Y)$ is true at the some model in $f(\neg X, EX)$, yet $f((\neg X \wedge Y) \vee (\neg X \wedge \neg Y), EX) \neq f(\neg X, EX)$. The selection function of CM does satisfy (S2), however; every model in $f(\phi, M)$ must be a model in which ϕ is true. It also satisfies (S3), on a natural understanding of “possible”; if there is a possible assignment of values to variables that makes A true, then $f(A, M) \neq \emptyset$.

5 Conclusion

A sufficiently developed causal modeling semantics for counterfactuals differs from the traditional similarity semantics—not just in the truth conditions it assigns to individual counterfactuals, but in the inferences it considers valid. How should proponents of causal modeling accounts interpret this fact? One possible conclusion is that causal models cannot provide a complete and correct account of the semantics of counterfactuals, and that instead, their function is primarily heuristic. Another possible conclusion is that causal models let us say all there is to be said about counterfactuals (at least, of the causal, non-backtracking sort), and we should revise our logic of counterfactuals.

Authors who adopt the second route have work to do, both technically and philosophically. Pertinent technical questions include: How should we extend the causal modeling semantics to incorporate quantifiers? And what axioms systems can be developed for other types of causal models—nonrecursive models, and models with indeterministic structural equations?

Pertinent philosophical questions are particularly acute in the philosophy of mind. Given the right sort of information about counterfactual dependence among neural states, and the supervenience of mental states on neural states, what sort of information can we derive about counterfactual dependence among mental states? Can the same event be counterfactually dependent both on a mental state and on the subvenient neural state?

The most interesting questions combine the technical and the philosophical. For instance: Can we make sense of such counterfactuals with counterfactual antecedents? On the present approach, the answer to this logical question turns on a series of metaphysical questions. Do counterfactuals have truthmakers? If so, how are these truthmakers related to the truthmakers for non-counterfactual sentences?

Another interesting set of technical-yet-philosophical questions involves *modus ponens*. Is it possible to somehow redefine the concept of an intervention in order to

rescue *modus ponens*? If so, is such a change desirable, in light of the conceptual role that interventions play?

Proponents of causal models need not answer any of these questions. Causal models are undoubtedly heuristically useful even if taken with a grain of metaphysical salt. But for those who choose to revise their logic, a fascinating research program awaits.

Acknowledgments I am thankful for the helpful questions and comments of audience members at the 2011 Carolina Metaphysics Workshop, the 2010 Formal Epistemology Workshop, and various workshops and departmental seminars in Australia. Special thanks are due to Delia Graff Fara and Peter Menzies for extensive commentary and discussion.

Appendix

Disjunctive Normal Form

Every Boolean sentence is exactly equivalent to a sentence in *disjunctive normal form*; i.e., a disjunction of conjunctions of atomic sentences. This disjunctive normal form can be derived in two steps.

The first step is to generate a sentence whose subsentences are all *positive*—i.e., all atomic sentences, negated atomic sentences, conjunctions, or disjunctions. Given a sentence ϕ_n with non-positive subsentences, one can generate a sentence ϕ_{n+1} by the following procedure. First, find ϕ_n 's longest non-positive subsentence ψ . (If two or more of these subsentences are tied for longest, let ψ be leftmost.)

If ψ is of the form $\neg\neg\theta$, then ϕ_{n+1} is the sentence that results from replacing ψ with θ in ϕ_n . (By $(\neg)$, this preserves exact equivalence.)

If ψ is of the form $\theta_1 \supset \theta_2$, then ϕ_{n+1} is the sentence that results from replacing ψ with $\neg\theta_1 \vee \theta_2$ in ϕ_n . (By $(\supset^+)$, this preserves exact equivalence.)

If ψ is of the form $\neg(\theta_1 \supset \theta_2)$, then ϕ_{n+1} is the sentence that results from replacing ψ with $\theta_1 \wedge \neg\theta_2$ in ϕ_n . (By $(\supset^-)$, this preserves exact equivalence.)

If ψ is of the form $\neg(\theta_1 \wedge \theta_2)$, then ϕ_{n+1} is the sentence that results from replacing ψ with $\neg\theta_1 \vee \neg\theta_2$ in ϕ_n . (By $(\wedge^-)$, this preserves exact equivalence.)

If ψ is of the form $\neg(\theta_1 \vee \theta_2)$, then ϕ_{n+1} is the sentence that results from replacing ψ with $\neg\theta_1 \wedge \neg\theta_2$ in ϕ_n . (By $(\vee^-)$, this preserves exact equivalence.)

Since this procedure replaces each non-positive subsentence of ϕ with a positive sentence, and generates either no new non-positive sentences, or one positive sentence which is shorter than the one replaced, it is guaranteed to terminate in a sentence with only positive subsentences. Since each step in the procedure preserves exact equivalence, the positive sentence derived at the end is exactly equivalent to the original sentence.

The next step is to “push the conjunctions inward”. Begin with a sentence ϕ_0 , all of whose subsentences are positive. Given a sentence ϕ_n , all of whose subsentences are positive, but not necessarily in disjunctive normal form, one can generate a sentence ϕ_{n+1} by the following procedure. First, find ϕ_n 's longest subsentence ψ that is a conjunction of disjunctions $(\theta_1^1 \vee \dots \vee \theta_{f(1)}^1) \vee \dots \vee (\theta_1^n \vee \dots \vee \theta_{f(n)}^n)$.

Consider all the conjunctions of θ_j^i s that have at least one conjunct of the form $\theta_{g(k)}^k$ for each $k \in \{1 \dots n\}$ (where $g(k)$ is an arbitrary number between 1 and $f(k)$ inclusive). Let ψ' be the disjunction of all these conjunctions. Let ϕ_{n+1} be the sentence that results from replacing ψ with the disjunction of all sentences of the form ψ' in ϕ_n . By $(\wedge^+)$ and $(\vee^+)$, replacing ϕ_n with ϕ_{n+1} preserves exact equivalence.

Each step of the second procedure eliminates the longest disjunction inside the scope of a conjunction, and without introducing any new disjunction inside the scope of a conjunction. Each step keeps the sentence positive. After a finite number of steps, the length of the longest disjunction inside the scope of a conjunction must equal 1, at which point $\phi = \phi_{n+1}$. At this stage, the sentence will be entirely positive, and have no disjunctions inside the scope of conjunctions: it will be in disjunctive normal form.

Soundness

The validity of (R1) and (CM1) is ensured by the definitions of the Boolean connectives.

(R2)'s validity is ensured by the selection semantics for $\Box \rightarrow$: Suppose $(\psi_1 \wedge \dots \wedge \psi_n) \supset \theta$ is true in every model, and $v_M((\phi \Box \rightarrow \psi_1) \wedge \dots \wedge (\phi \Box \rightarrow \psi_n)) = 1$. It remains to show that $v_M(\phi \Box \rightarrow \theta) = 1$. The semantic rule for $\wedge$ ensures that for each of $\psi_i \in \{\psi_1 \dots \psi_n\}$, $v_M(\psi_i) = 1$. By the semantic rule for $\Box \rightarrow$, for every $\psi_i \in \{\psi_1 \dots \psi_n\}$ and $M' \in f(\phi, M)$, $v_{M'}(\psi_i) = 1$. By hypothesis, $v_{M'}((\psi_1 \wedge \dots \wedge \psi_n) \supset \theta) = 1$ for every $M' \in f(\phi, M)$. By tautological reasoning $v_{M'}(\theta) = 1$ for every $M' \in f(\phi, M)$, and so $v_M(\phi \Box \rightarrow \theta) = 1$.

The validity of (CM2) and (CM3) is ensured by the behaviour of variables. Every variable takes on some value or other (hence (CM2)), and no variable takes on two values at once (hence (CM3)).

(CM4) is valid because every sentence of the form **A** is made true by a single state, and each state corresponds to exactly one intervention. So $f(\mathbf{A}, M)$ contains exactly one model M' . For every sentence ϕ , either $v_{M'}(\theta) = 1$, in which case $v_M(\mathbf{A} \Box \rightarrow \phi) = 1$, or $v_{M'}(\theta) = 0$, in which case $v_M(\mathbf{A} \Box \rightarrow \neg \phi) = 1$. In either case $v_M((\mathbf{A} \Box \rightarrow \phi) \vee (\mathbf{A} \Box \rightarrow \neg \phi)) = 1$.

(CM5) is valid because every sentence of the form **A** is made true by some state or other, and each state corresponds to an intervention. Therefore, there is at least one submodel in $f(\mathbf{A}, M)$. Since $\perp$ is true in no model, $\perp$ is not true in every submodel in $f(\mathbf{A}, M)$. Hence, $v_M(\neg(\mathbf{A} \Box \rightarrow \perp)) = 1$.

(CM6) is valid because the order of interventions on a causal model does not affect the end result of those interventions, provided one doesn't intervene twice on the same variable. To generate $M_{\mathbf{A}} \wedge V = v$ from $M_{\mathbf{A}}$, one intervenes on $M_{\mathbf{A}}$ to set $V = v$. Suppose $v_{M_{\mathbf{A}}}(V = v) = 0$. Then by the semantic rules for material conditional and counterfactual, $v_M(\mathbf{A} \Box \rightarrow (V = v \supset B)) = 1$. By tautological reasoning $v_M(((\mathbf{A} \wedge V = v) \Box \rightarrow B) \supset (\mathbf{A} \Box \rightarrow (V = v \supset B))) = 1$. On the other hand, suppose that in $M_{\mathbf{A}}$, $V = v$. Intervening to change V 's structural equation will not affect the values of V 's nondescendants, for their values do not depend on V 's

value. Nor (by hypothesis) will it affect V 's value. Nor will it affect the values of V 's descendants, since in a recursive model, their values are determined by the value of V and V 's nondescendants. Ergo, $v_{M_{\mathbf{A}}}(V' = v') = v_{M_{\mathbf{A}}} \wedge V = v(V' = v')$ for all $V' \in \mathcal{V}$. Hence, where B is any Boolean combination of atomic sentences, $v_{M_{\mathbf{A}}}(B) = v_{M_{\mathbf{A}}} \wedge V = v(B)$, and so once again, $v_M(((\mathbf{A} \wedge V = v) \Box \rightarrow B) \supset (\mathbf{A} \Box \rightarrow (V = v \supset B))) = 1$.

(CM7) is valid because any intervention that makes A true also generates a model in which A is true. This holds where A is an atomic sentence—intervening to set $V = v$ generates a model in which $V = v$ —or a negated atomic sentence—intervening to set $V = v'$ generates a model in which $V \neq v$, so long as $v' \neq v$. The conditions on truthmaking ensure that if it holds for ϕ and ψ , it also holds for $\phi \vee \psi$, $\neg(\phi \vee \psi)$, $\phi \wedge \psi$, $\neg(\phi \wedge \psi)$, $\phi \supset \psi$ and $\neg(\phi \supset \psi)$.

(CM8) is valid for the class of recursive models. To show this, we must appeal to a notion of counterfactual dependence among variables. Say that V' *counterfactually depends* on V iff for some variables $V_1 \dots V_m$, all distinct from V and V' , some values $v_1 \dots v_m$ of $V_1 \dots V_m$, some values v and v^* of V , and some value v' of V' , the intervention setting $V_1 = v_1, \dots V_m = v_m$, and $V = v$ generates a submodel in which $V' = v'$, while the intervention setting $V_1 = v_1, \dots V_m = v_m$, and $V = v^*$ generates a submodel in which $V' \neq v'$.

We can show that V' stands in the ancestral of counterfactual dependence to V iff V is an ancestor of V' . First, the left-to-right direction: If V' counterfactually depends on V , then the structural equations ensure that V' 's value depends non-trivially on V 's value. Hence, V is an ancestor of V' . Since the relation of being an ancestor is transitive, it follows that if V' stands in the ancestral of counterfactual dependence to V , then V is an ancestor of V' . Next, the right-to-left direction: if V_1 is an ancestor of V_n , there is a sequence of variables $V_1, \dots V_n$ such that V_1 is the parent of V_2 , and $\dots V_{n-1}$ is the parent of V_n . Since each variable counterfactually depends on each of its parents, it follows that V' stands in the ancestral of counterfactual dependence to V .

Furthermore, it follows from the definition of the selection function and the semantic rule for the counterfactual that V' counterfactually depends on V iff there is a sentence of the form $d(V, V')$ such that $v_M(d(V, V')) = 1$. As a corollary, if there exist sentences $d_1(V_1, V_2) \dots d_{n-2}(V_{n-2}, V_{n-1})$ such that $v_M(d_1(V_1, V_2) \wedge \dots \wedge d_{n-1}(V_{n-1}, V_n)) = 1$, then V_n stands in the ancestral of counterfactual dependence to V_1 , and so V_1 is an ancestor of V_n .

In order to show the validity of (CM8), we can assume that $v_M(d_1(V_1, V_2) \wedge \dots \wedge d_{n-1}(V_{n-1}, V_n)) = 1$, and show that there is no sentence of the form $d_n(V_n, V_1)$ such that $v_M(d_n(V_n, V_1)) = 1$. By our supposition, in M , V_n stands in the ancestral of the counterfactual dependence relation to V_1 . Therefore, V_1 is an ancestor of V_n . On the assumption that M is recursive, V_{n-1} is not an ancestor of V_1 , and so cannot counterfactually depend on V_{n-1} . Therefore, no sentence of the form $d_n(V_{n-1}, V_1)$ can be true in M ; QED.

(CM9) is sound according to the recursive conditions for truthmaking. $v_M((A \vee B) \Box \rightarrow \phi) = 1$ iff $v_{M'}(\phi) = 1$ for every submodel M' generated by an intervention on M corresponding to a state that makes $A \vee B$ true. But according to condition $(\vee^+)$, $s \models A \vee B$ iff $s \models A$ or $s \models B$ or $s = t \sqcup r$ where $t \models A$ and $r \models B$. In other words,

rephrasing the third disjunct in terms of condition $(\wedge^+)$, $s \models A \vee B$ iff $s \models A$ or $s \models B$ or $s \vdash A \wedge B$. So $v_M((A \vee B) \Box \rightarrow \phi) = 1$ iff $v_{M'}(\phi) = 1$ for every submodel M' generated by an intervention on M corresponding to a state that makes A true, and $v_{M'}(\phi) = 1$ for every submodel M' generated by an intervention on M corresponding to a state that makes B true, and $v_{M'}(\phi) = 1$ for every submodel M' generated by an intervention on M corresponding to a state that makes $A \wedge B$ true. In other words, $v_M((A \vee B) \Box \rightarrow \phi) = 1$ iff $v_M(A \Box \rightarrow \phi) = 1$, and $v_M(B \Box \rightarrow \phi) = 1$, and $v_M((A \wedge B) \Box \rightarrow \phi) = 1$. The semantic rules for Boolean connectives ensure that $v_M(((A \vee B) \Box \rightarrow \phi) \equiv ((A \Box \rightarrow \phi) \wedge (B \Box \rightarrow \phi) \wedge ((A \wedge B) \Box \rightarrow \phi))) = 1$.

(CM10) is valid because any exact equivalent A' of A is made true by the same states as A ; hence $f(A, M) = f(A', M)$.

(CM11) is valid because where $\mathcal{V}_A$ is the set of variables mentioned in A , $\mathcal{V}_B$ is the set of variables mentioned in B and B' , and $\mathcal{V}_C$ is the set of variables mentioned in C , the procedure of performing the interventions sanctioned by each of A , B , and C has the same results as the procedure by performing the interventions sanctioned by A and B' , then performing the interventions sanctioned by B (thereby reversing the interventions sanctioned by B') and performing the interventions sanctioned by C .

Completeness

A proof system is complete over a class of models iff for every consistent set of sentences Γ , there is a model in which every sentence in Γ is true. Suppose, then, that we have a consistent set of sentences Γ_0 . If we can specify a procedure for constructing a model in which every member of Γ_0 is true, then we will have proved completeness. We begin by using Lindenbaum's Lemma to construct a maximal consistent set Γ such that $\Gamma_0 \subseteq \Gamma$. We can use Γ to build a causal model $M = \langle \mathcal{V}, \mathcal{S}, \mathcal{A} \rangle$, by the following procedure.

Let $\mathcal{V}$ contain one variable V_i for each variable V_i for which the language contains a name.

To define $\mathcal{S}$, we must show that for every $V_i \in \mathcal{V}$ and partial assignment function $\mathcal{A}^-$, where $\mathcal{A}^-(V_j) \in \mathcal{R}(V_j)$ for the $V_j \in \mathcal{V} - V_i$, Γ contains exactly one sentence of the form $\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i = v_i$. To show this, we must prove both an existence claim and a uniqueness claim.

The existence claim: for every $V_i \in \mathcal{V}$ and partial assignment function $\mathcal{A}^-$, Γ contains at least one sentence of the form $\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i = v_i$. This can be shown by assuming otherwise, and proving a contradiction. Since Γ is maximal, if the existence claim is false, then for every $v_i \in \mathcal{R}(V_i)$,

$$\neg(\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i = v_i) \in \Gamma$$

Since Γ contains every instance of every axiom schema, and (CM4) is an axiom, for each $v_i \in \mathcal{R}(V_i)$,

$$(\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i = v_i) \vee (\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i \neq v_i) \in \Gamma$$

Since Γ is closed under tautological consequence, for each $v_i \in \mathcal{R}(V_i)$,

$$\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i \neq v_i \in \Gamma$$

By (R2) and the closure of Γ under logical consequence,

$$\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow \bigwedge_{v \in \mathcal{R}(V_i)} V_i \neq v \in \Gamma$$

By (CM1), (CM2), and (R1),

$$\bigwedge_{v \in \mathcal{R}(V_i)} V_i \neq v \vdash \perp$$

and so by (R2), and the closure of Γ under logical consequence,

$$\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow \perp \in \Gamma$$

But since every instance of (CM5) is an axiom,

$$\neg(\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow \perp) \in \Gamma$$

We have derived a contradiction, thus completing the existence proof.

Now, the uniqueness claim: for every $V_i \in \mathcal{V}$ and partial assignment function $\mathcal{A}^-$, Γ contains at most one sentence of the form $\bigwedge_{j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i = v_i$. Again, this can be shown by assuming otherwise and proving a contradiction. We will suppose the uniqueness claim is false, i.e., that Γ contains sentences of the form

$$\bigwedge_{j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i = v_i$$

and

$$\bigwedge_{j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow V_i = v'_i$$

where $v_i \neq v'_i$. And we will prove a contradiction.

By (R2) and the closure of Γ under entailment, we can infer that

$$\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow (V_i = v_i \wedge V_i = v'_i) \in \Gamma$$

But by (CM3)

$$V_i = v_i \wedge V_i = v'_i \vdash \perp$$

and so by (R2),

$$\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow \perp \in \Gamma$$

But since every instance of (CM5) is an axiom,

$$\neg(\bigwedge_{V_j \in \mathcal{V} - V_i} (V_j = \mathcal{A}^-(V_j)) \Box \rightarrow \perp) \in \Gamma$$

We have derived a contradiction, thus completing the uniqueness proof.

Now we can define $\mathcal{S}$. Since there are many possible partial assignment functions for each set of variables, let us number each possible $\mathcal{A}^-$ with a subscript. First, let us write, for each V_i , an equation of the form

$$f_i = v_k \begin{array}{l} \text{if } (V_j = \mathcal{A}_k^-(j)) \text{ for each } V_j \in \mathcal{V} - \{V_i\} \\ \text{where } \bigwedge_{j \in \mathcal{V} - V_i} (V_j = \mathcal{A}_k^-(j)) \Box \rightarrow V_i = v_k \in \Gamma \end{array}$$

By our existence and uniqueness claim, there is exactly one such v_k for every i and k . Now, let us replace each of these equations with an equation that has the same

solutions in all cases, but does not mention any redundant variables. (If all variables are redundant, delete the equation altogether.) The set of equations generated this way is $\mathcal{S}$. $\mathcal{S}$ will contain at most one structural equation for each variable.

The last step in defining M is to define $\mathcal{A}$. (CM2) and (CM3) ensure that for each $V \in \mathcal{V}$, Γ contains exactly one sentence of the form $V = v$. Let $\mathcal{A}(V_i)$ be the v_i such that $V_i = v_i \in \Gamma$.

Having defined M , we must show that it is recursive. (CM8) ensures that for any two variables V and V' , if V' stands in the ancestral of counterfactual dependence to V , then V does not counterfactually depend on V' . I have shown in the soundness portion of the appendix that the ancestral of counterfactual dependence is the ancestor relation. It follows that for any two variables V and V' , if V' is an ancestor of V , then V is not counterfactually dependent on V' . Furthermore, since all variables counterfactually depend on their parents, we can infer that for any two variables V and V' , if V' is an ancestor of V , then V is not parent of V' . This entails that no variable is its own descendant. For if a variable V were its own descendant, it would have a parent V' that was also V 's descendant. Thus, no variable in M can be its own descendant, and so M is recursive.

The last step of the completeness proof is to show that if $\phi \in \Gamma$, then $v_M(\phi) = 1$. This can be proved in two steps. First, we prove that every sentence is equivalent to a Boolean compound of atomic sentences and simple conditionals of the form $\mathbf{A} \square \rightarrow B$. Next, we show that for any such Boolean compound, $\phi \in \Gamma$ iff $v_M(\phi) = 1$.

I will prove that every sentence is equivalent to a Boolean compound of atomic sentences and simple conditionals of the form $\mathbf{A} \square \rightarrow B$ by providing a translation procedure.

The first part of the translation procedure eliminates all complex antecedents. Begin with a sentence ϕ_0 . Let ψ_n be ϕ_n 's shortest subsentence of the form $\mathbf{A} \square \rightarrow \theta$. By (CM10) and the disjunctive normal form theorem, θ is equivalent to a sentence of the form $\mathbf{A}_1 \vee \dots \vee \mathbf{A}_n \square \rightarrow \psi$. By (CM9), ψ_n is equivalent to a conjunction of the form $(\mathbf{A}_1 \square \rightarrow \psi) \wedge \dots \wedge (\mathbf{A}_n \square \rightarrow \psi)$. Let ϕ_{n+1} be the result of replacing ψ with this equivalent sentence. Repeated applications of (R2) ensure ϕ_{n+1} is equivalent to ϕ_n . Each step eliminates a counterfactual with a complex antecedent without setting up a new one in its place, and so the procedure will terminate when there are no remaining subsentences with complex antecedents.

The second part of the translation procedure is to eliminate counterfactual consequents. Begin with a sentence ϕ_0 , all of whose conditional subsentences have antecedents of the form $\mathbf{A}$. Let ψ_n be ϕ_n 's shortest subsentence of the form $\mathbf{A} \square \rightarrow \theta$ where θ is not of the form B . Let $(\theta_1^1 \wedge \dots \wedge \theta_1^{f(1)}) \vee \dots \vee (\theta_n^1 \wedge \dots \wedge \theta_n^{f(n)})$ be the (classical) disjunctive normal form of θ . By (R2), ψ' is equivalent to $\mathbf{A} \square \rightarrow ((\theta_1^1 \wedge \dots \wedge \theta_1^{f(1)}) \vee \dots \vee (\theta_n^1 \wedge \dots \wedge \theta_n^{f(n)}))$, and by (CM4) and (R2) together, ϕ' is equivalent to $((\mathbf{A} \square \rightarrow \theta_1^1) \wedge \dots \wedge (\mathbf{A} \square \rightarrow \theta_1^{f(1)})) \vee \dots \vee ((\mathbf{A} \square \rightarrow \theta_n^1) \wedge \dots \wedge (\mathbf{A} \square \rightarrow \theta_n^{f(n)}))$. Each θ_j^i is either atomic or of the form $\mathbf{B} \square \rightarrow C$. If θ_j^i is of the form $\mathbf{B}_i^j \square \rightarrow C_i^j$, then (CM11) ensures that $\mathbf{A} \square \rightarrow \theta_j^i$ is equivalent to $\mathbf{A}_i^j \wedge \mathbf{B}_i^j \square \rightarrow C_i^j$, where $\mathbf{A}_i^j$ the conjunction that results from removing any conjuncts from $\mathbf{A}$ that

share variables with $\mathbf{B}_i^j$. Finally, let $\mathbf{A}_i^j = A$ and $C_i^j = \theta_i^j$ if θ_i^j is atomic. Thus, ψ_n is equivalent to $((\mathbf{A}_1^1 \Box \rightarrow C_1^1) \wedge \dots \wedge (\mathbf{A}_1^{f(1)} \Box \rightarrow C_1^{f(1)})) \vee \dots \vee ((\mathbf{A}_n^1 \Box \rightarrow C_n^1) \wedge \dots \wedge (\mathbf{A}_n^{f(n)} \Box \rightarrow C_n^{f(n)}))$ (call this sentence ψ'_n). Let ϕ_{n+1} be the sentence that results from replacing ψ_n with ψ'_n in ϕ_n . Repeated applications of (R2) ensure that ϕ_{n+1} is equivalent to ϕ_n . Each step eliminates a counterfactual consequent without setting up a new one in its place, and so the procedure will terminate when there are no remaining subsentences with counterfactual consequents.

All that remains now is to show that if ϕ is a Boolean compound of atomic sentences and sentences of the form $\mathbf{A} \Box \rightarrow B$, then $v_M(\phi) = 1$. In the base case, where ϕ is an atomic sentence, this follows straightforwardly from the definition of $\mathcal{A}$. Suppose it holds for all sentences shorter than ϕ . Then we can show that it holds for ϕ as well.

- Suppose ϕ is of the form $V \neq v$ and $\phi \in \Gamma$. We have already shown that if $\phi \in \Gamma$, there is some $v' \neq v$ such that $V = v' \in \Gamma$. By the definition of $\mathcal{A}$, $\mathcal{A}(V) = v'$. By the semantic rule for atomic sentences $v_M(V = v) = 0$, and so by the semantic rule for negation, $v_M(\phi) = 1$.
- Where ϕ is of the form $\psi_1 \vee \psi_2, \neg(\psi_1 \vee \psi_2), \psi_1 \wedge \psi_2, \neg(\psi_1 \wedge \psi_2), \psi_1 \supset \psi_2$, or $\neg(\psi_1 \supset \psi_2)$, we can use the usual definitions of truth functional connectives to prove on the basis of the inductive hypothesis that $\phi \in \Gamma$.
- Suppose ϕ is of the form $\mathbf{A} \Box \rightarrow V = v$ and $\phi \in \Gamma$. We can prove by induction on the number of variables in $\mathbf{A}$ (written ' $|\mathbf{A}|$ ') that $v_M(\phi) = 1$. (CM7) and (CM5) ensure that this hypothesis holds for the case of $|\mathbf{A}| = |\mathcal{V}|$; the definition of the selection function ensures that it holds where $|\mathbf{A}| = |\mathcal{V}| - 1$. Suppose it holds for $|\mathbf{A}| > |\mathcal{V}| - n$. We can show that it holds for $|\mathbf{A}| = |\mathcal{V}| - n$. For any two distinct variables V and V' not mentioned in $\mathbf{A}$, there will be a v and a v' such that Γ contains sentences of the form $(\mathbf{A} \wedge V' = v') \Box \rightarrow V = v$ and $(\mathbf{A} \wedge V = v) \Box \rightarrow V' = v'$.
- (We know there must be such sentences by (CM4), (CM2), (R2), and (CM8). (CM4), (CM2), and (R2) ensure that Γ contains sentences of the form $(\mathbf{A} \wedge V' = v') \Box \rightarrow V = v_1$ and $(\mathbf{A} \wedge V = v) \Box \rightarrow V' = v'_1$ for each $v \in \mathcal{R}(V)$ and $v' \in \mathcal{R}(V')$. Since M is recursive, one of the variables is a nondescendant of the other. Let us assume without loss of generality that V is a nondescendant of V' . There must be a v such that $(\mathbf{A} \wedge V' = v'_i) \Box \rightarrow V = v \in \Gamma$ for every $v'_i \in \mathcal{R}(V')$. We can simply choose for our v' the $v \in \mathcal{R}(V)$ such that $(\mathbf{A} \wedge V = v) \Box \rightarrow V' = v'$.)
- By the inductive hypothesis, $v_M((\mathbf{A} \wedge V' = v') \Box \rightarrow V = v) = v_M((\mathbf{A} \wedge V = v) \Box \rightarrow V' = v') = 1$. Since (we've assumed without loss of generality) V is a nondescendant of V' , $(\mathbf{A} \wedge V' = v'_i) \Box \rightarrow V = v \in \Gamma$ for every $v'_i \in \mathcal{R}(V')$. By (CM6), $\mathbf{A} \Box \rightarrow (V' = v'_i \supset V = v) \in \Gamma$ for every $v'_i \in \mathcal{R}(V')$. By (R2) and (CM2), $\mathbf{A} \Box \rightarrow V = v \in \Gamma$.
- We know that $v_M(\mathbf{A} \Box \rightarrow V = v) = 1$, since $\mathbf{A} \Box \rightarrow V = v$ follows validly from other sentences that are true in M , and all our axioms and rules are sound. (R2), (CM2), and (CM5) ensure that Γ contains no sentence of the form $\mathbf{A} \Box \rightarrow V = v^*$ for $v^* \neq v$.
- Thus, for arbitrary V and $v \in \mathcal{R}(V)$, if $|\mathbf{A}| \leq |\mathcal{V}| - (n + 1)$, if $\mathbf{A} \Box \rightarrow V = v \in \Gamma$, then $v_M(\mathbf{A} \Box \rightarrow V = v) = 1$.

- Suppose ϕ is of the form $A \Box \rightarrow B$. (CM4) ensures that for every atomic subsentence $V_i = v_i$ of B , either $A \Box \rightarrow V_i = v_i \in \Gamma$ or $A \Box \rightarrow V_i \neq v_i \in \Gamma$. (R2) and (CM5) entail that the conjunction of the $V_i = v_i \in \Gamma$ and the $V_i \neq v_i \in \Gamma$ is compatible with B ; therefore this conjunction entails B . By our inductive hypothesis, $v_M(A \Box \rightarrow V_i = v_i) = 1$ for all V_i such that $A \Box \rightarrow V_i = v_i \in \Gamma$, and $v_M(A \Box \rightarrow V_i \neq v_i) = 1$ for all V_i such that $A \Box \rightarrow V_i \neq v_i \in \Gamma$. By the selection semantics for the counterfactual, $v_M(A \Box \rightarrow B) = 1$.

References

- Edgington, D. (2004). Counterfactuals and the benefit of hindsight. In: P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world*. Routledge.
- Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science*, 68S, S313–S324.
- Fine, K. (1975). Critical notice. *Mind*, 84(1), 451–458.
- Fine, K. (forthcoming). Counterfactuals without possible worlds. *The Journal of Philosophy*.
- Fodor, J. (1974). Special sciences, or the disunity of science as a working hypothesis. *Synthese*, 28, 97–115.
- Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1), 151–182.
- Gundersen, L. B. (2004). Outline of a new semantics for counterfactuals. *Pacific Philosophical Quarterly*, 85(1), 1–20.
- Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, 132(1), 109–136.
- Halpern, J. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12, 317–337.
- Handfield, T., Twardy, C. R., Korb, K. B., & Oppy, G. (2008). The metaphysics of causal models: Where's the biff? *Erkenntnis*, 68(2), 149–68.
- Hawthorne, J. (2005) Chance and counterfactuals. *Philosophy and Phenomenological Research*, 70(2), 396–405.
- Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous*, 39(4), 232–257.
- Hitchcock, C. (2001). The intransitivity of causation revealed in arrows and graphs. *The Journal of Philosophy*, 98(6), 273–299.
- Hüber, F. Structural equations and beyond. manuscript.
- Kment, B. (2006). Counterfactuals and explanation. *Mind*, 115(458), 261–310.
- Krasner, D., & Heller, M. (1994). The miracle of counterfactuals: Counterexamples to lewis's world ordering. *Philosophical Studies*, 76(1), 27–43.
- Lewis, D. (1973a). Causation. *The Journal of Philosophy*, 70(17), 556–567.
- Lewis, D. K. (1973b). *Counterfactuals*. Harvard University Press.
- Lewis, D. (1979) Counterfactual dependence and time's arrow. *Nous*, 13(4), 455–476.
- Lewis, D. (1981). Counterfactuals and comparative possibility. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs, volume 15 of The University of Western Ontario series in philosophy of science* (pp. 57–85). Dordrecht: D. Reidel.
- Lewis, D. (1986a). Introduction. In: *Philosophical papers: Volume II*. New York: Oxford University Press.
- Lewis, D. (1986b). A subjectivist's guide to objective chance. In *Philosophical papers* (Vol. II). New York: Oxford University Press.
- Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103, 473–490.
- Loewer, B. (1976). Counterfactuals with disjunctive antecedents. *The Journal of Philosophy*, 73(16), 531–537.
- McGee, V. (1985). A counterexample to modus ponens. *The Journal of Philosophy*, 82(9), 462–471.
- Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science*, 71(5):820–832.

- Nute, D. (1978). Simplification and substitution of counterfactual antecedents. *Philosophia*, 7(2), 317–325.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: University Press.
- Penczek, A. (1997). Counterfactuals with true components. *Erkenntnis*, 46(1), 79–85.
- Sartorio, C. (2006). Disjunctive causes. *Journal of Philosophy*, 103
- Schaffer, J. (2004). Counterfactuals, causal independence and conceptual circularity. *Logic and Analysis*, 64(4), 299–308.
- Stalnaker, R. (1981a). A defense of conditional excluded middle. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs, volume 15 of The University of Western Ontario series in philosophy of science* (pp. 87–104). Dordrecht: D. Reidel.
- Stalnaker, R. (1981b). A theory of conditionals. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs, volume 15 of The University of Western Ontario series in philosophy of science* (pp. 41–55). Dordrecht: D. Reidel.
- Wasserman, R. (2006). The future similarity objection revisited. *Synthese*, 150(1), 57–67.
- Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.
- Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part I: A counterfactual account. *Nóûs*, 37(1), 1–24.