THE REVIEW OF SYMBOLIC LOGIC
Volume 6, Number 2, June 2013

FROM CAUSAL MODELS TO COUNTERFACTUAL STRUCTURES

JOSEPH Y. HALPERN
Computer Science Department, Cornell University

Abstract. Galles & Pearl (1998) claimed that "for recursive models, the causal model framework does not add any restrictions to counterfactuals, beyond those imposed by Lewis's [possible-worlds] framework." This claim is examined carefully, with the goal of clarifying the exact relationship between causal models and Lewis's framework. Recursive models are shown to correspond precisely to a subclass of (possible-world) counterfactual structures. On the other hand, a slight generalization of recursive models, models where all equations have unique solutions, is shown to be incomparable in expressive power to counterfactual structures, despite the fact that the Galles and Pearl arguments should apply to them as well. The problem with the Galles and Pearl argument is identified: an axiom that they viewed as irrelevant, because it involved disjunction (which was not in their language), is not irrelevant at all.

§1. Introduction. Counterfactual reasoning arises in broad array of fields, from statistics to economics to law. Not surprisingly, there has been a great deal of work on giving semantics to counterfactuals. Perhaps the best-known approach is due to Lewis (1973) and Stalnaker (1968), and involves possible worlds. The idea is that a counterfactual of the form "if A were the case then B would be the case," typically written $A \succeq B$, is true at a world w if B is true at all the worlds closest to w where A is true. Of course, making this precise requires having some notion of "closeness" among worlds.

More recently, Pearl (2000) proposed the use of causal models based on *structural equations* for reasoning about causality. In causal models, we can examine the effect of *interventions*, and answer questions of the form "if random variable X were set to x, what would the value of random variable Y be." This suggests that causal models can also provide semantics for (at least some) counterfactuals.

The relationship between the semantics of counterfactuals in causal models and in counterfactual structures (i.e., possible-worlds structures where the semantics of counterfactuals is given in terms of closest worlds) was studied by Galles & Pearl (1998). They argue that the relationship between the two approaches depends in part on whether we consider *recursive* (i.e., *acyclic*) models (those without feedback—see Section §2 for details). They reach the following conclusion (Pearl, 2000, p. 242).[1]

> In sum, for recursive models, the causal model framework does not add any restrictions to counterfactuals beyond those imposed by Lewis's framework; the very general concept of closest worlds is sufficient. Put another way, the assumption of recursiveness is so strong that it already embodies all other restrictions imposed by structural semantics. When

Received: April 14, 2012.

[1] The discussion in Section 7.4.2 of Pearl (2000) is taken almost verbatim from Galles & Pearl (1998); since the former is more widely available, that is what I cite in this paper.

© Association for Symbolic Logic, 2012
doi:10.1017/S1755020312000305

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

> we consider nonrecursive systems, however, we see that reversibility [a particular axiom introduced by Galles and Pearl] is not enforced by Lewis's framework.

This conclusion seems to have been accepted by the community. For example, in the Wikipedia article on "Counterfactual conditional" (http://en.wikipedia.org/wiki/Counterfactual_conditional; September 2009), it says "[The definition of counterfactuals in causal models] has been shown to be compatible with the axioms of possible world semantics."

In this paper I examine these claims and the proofs given for them more carefully, and try to settle once and for all the relationship between causal models and counterfactual structures. The first sentence in the statement above says "for recursive models, the causal model framework does not add any restrictions to counterfactuals beyond those imposed by Lewis's framework." It is not clear (at least to me) exactly what this means. Recursive models are a well-defined subclass of causal models. Galles and Pearl themselves show that there are additional axioms that hold in recursive models over and above those that hold in counterfactual structures. Indeed, they show that the reversibility axiom mentioned above is valid in recursive models and is not valid in possible-worlds models. They also show that all the axioms of causal reasoning in the possible-worlds framework that they view as relevant (specifically, axioms that do not mention disjunction, since it is not in their language) hold in recursive causal models. Thus, the only conclusion that can be drawn from their argument is just the opposite to what they claim, namely, that the possible-worlds approach does not add any restrictions to causal reasoning beyond those imposed by causal models, since causal models satisfy all the axioms that counterfactual structures do, and perhaps more.[2]

Pearl (private communication, 2010) intended the clause "for recursive models" to apply to counterfactual structures as well as to structural models. However, the notion of a recursive counterfactual structure is not defined either by Galles & Pearl (1998) or Pearl (2000). In fact, in general, the notion of recursive model as defined, for example, in Pearl (2000, Definition 7.3.4), does not even make sense for counterfactual structures. I show that, if we focus on the language considered by Galles and Pearl and counterfactual structures appropriate for this language, there is a well-defined subclass of counterfactual structures that can justifiably be viewed as recursive (counterfactual) structures. I then show that precisely the same formulas in this language are valid in recursive causal models and recursive counterfactual structures. Put another way, at least as far as Galles and Pearl's language goes, recursive causal models and recursive counterfactual structures are equally expressive. Thus, by restricting to recursive counterfactual structures (as Pearl intended), the Galles–Pearl claim is, in a sense, correct (although the claim does not follow from their argument—see below). However, it should be noted that Galles and Pearl's language cannot, for example, express statements with disjunctive antecedents, such as "if he had chosen a different running mate or had spent his campaign funding more wisely then he would have won the election." The claim applies only to their restricted language.

[2] Although it is not relevant to the focus of this paper, I in fact also do not understand the second sentence in the Pearl quote above. As for the third sentence, while it is the case that reversibility does not hold in counterfactual structures in general, reversibility holds in recursive causal structures as well as more general causal structures, as Pearl and Galles themselves show.

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

Galles and Pearl try to prove their claims by considering axioms; my proof works at the level of structures. Specifically, I show that a recursive causal model satisfying a particular formula can be effectively converted to a recursive counterfactual structures satisfying this formula, and vice versa. It is actually important to work at the level of structures to prove this claim, rather than working at the level of axioms. I show that the argument that Galles and Pearl give for their claim is incorrect in some significant respects. The problem lies with their claim that axioms involving disjunctive antecedents are irrelevant. As I show, they are quite relevant; a proof of a formula not involving disjunction in the antecedent may need to use an axiom that does. This possibility is illustrated by considering a class of models slightly larger than recursive models. In a recursive model, given a context, there is a unique solution to every equation. In earlier work (Halpern, 2000), I showed that there are nonrecursive causal models where every equation has a unique solution. Galles and Pearl's argument applies without change to causal models where the equations have a unique solution. However, as I show here, these models are actually *incomparable* in expressive power to counterfactual structures. The reversibility axiom remains valid in causal models where the equations have a unique solution. But, as I show by example, there is a formula that does not involve disjunction that is valid in counterfactual structures but is *not* valid in causal models where the equations have a unique solution. Not surprisingly, proving that this formula is valid from the axioms requires the use of an axiom that involves disjunction.

These results show that the quote from Wikipedia is not quite true. While it is true that *recursive* causal structures are, in a certain sense, compatible with the axioms of possible worlds semantics, the slightly more general class of causal structures with unique solutions to the equations is not. Thus, in general, the semantics of counterfactuals in causal structures *cannot* be understood in terms of closest worlds.

The rest of this paper is organized as follows. In Section §2, I review the causal models and counterfactual structures. The main technical results are in Section §3, which also includes a discussion of the problems with the Galles–Pearl argument. I conclude in Section §4 with some discussion of these results.

§2. Causal models and possible-worlds models: a review. In this section I review the relevant material on causal models and possible-worlds models.

2.1. Causal models. (The following discussion is taken, with minor modifications, from Halpern, 2000.) Causal models describe the world in terms of random variables, some of which have a causal effect on others. It is conceptually useful to split the random variables into two sets, the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. The values of the endogenous variables are characterized by a set of *structural equations*.

For example, if we are trying to determine whether a forest fire was caused by lightning or an arsonist, we can take the world to be described by four random variables:

- FF for forest fire, where $FF = 1$ if there is a forest fire and $FF = 0$ otherwise;
- L for lightning, where $L = 1$ if lightning occurred and $L = 0$ otherwise;
- MD for match dropped (by arsonist), where $MD = 1$ if the arsonist dropped a lit match, and $MD = 0$ otherwise;
- E, which captures the external factors that determine whether the arsonist will drop a match, or whether lightning will strike.

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

The variables FF, L, and MD are endogenous, while E is exogenous. If we want to model the fact that if the arsonist drops a match *or* lightning strikes then a fire starts, then we would have the equation $FF = \max(L, MD)$; that is, the value of the random variable FF is the maximum of the values of the random variables MD and L. This equation says, among other things, that if $MD = 0$ and $L = 1$, then $FF = 1$. Alternatively, if we want to model the fact that a fire requires both a lightning strike *and* a dropped match (perhaps the wood is so wet that it needs two sources of fire to get going), then the only change in the model is that the equation for FF becomes $FF = \min(L, MD)$; the value of FF is the minimum of the values of MD and L. The only way that $FF = 1$ is if both $L = 1$ and $MD = 1$.

Formally, a *signature* $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V}$ is a finite set of endogenous variables, and $\mathcal{R}$ associates with every variable $X \in \mathcal{U} \cup \mathcal{V}$ a finite set $\mathcal{R}(X)$ of possible values for X (the *range* of possible values of X). A *causal model* is a pair $T = (\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a signature and $\mathcal{F}$ associates with each variable $X \in \mathcal{V}$ a function denoted F_X such that $F_X : (\times_{Z \in (\mathcal{U} \cup \mathcal{V} - \{X\})} \mathcal{R}(Z)) \rightarrow \mathcal{R}(X)$. F_X characterizes the value of X given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. Because F_X is a function, there is a unique value of X once all the other variables are set. Notice that we have such functions only for the endogenous variables. The exogenous variables are taken as given; it is their effect on the endogenous variables (and the effect of the endogenous variables on each other) that is modeled by the structural equations.

Given a causal model $T = (\mathcal{S}, \mathcal{F})$, a (possibly empty) vector $\vec{X}$ of variables in $\mathcal{V}$, and a vector $\vec{x}$ of values for the variables in $\vec{X}$, we can define a new causal model denoted $T_{\vec{X}=\vec{x}}$. Intuitively, this is the causal model that results when the variables in $\vec{X}$ are set to $\vec{x}$. We can think of setting $\vec{X}$ to $\vec{x}$ as an intervention. For example, if T is the causal model for the forest fire described above, where $FF = \max(L, MD)$, then $T_{L=0}$ is the model where the lightning definitely does not occur, so that there is a forest fire if and only if the arsonist drops a match. If T' is the model where $FF = \min(L, MD)$, then $T'_{L=0}$ is the model where there is no forest fire, since there is no lightning.

Formally, $T_{\vec{X}=\vec{x}} = (\mathcal{S}, \mathcal{F}^{\vec{X}=\vec{x}})$, where $\mathcal{F}^{\vec{X}=\vec{x}}$ is identical to $\mathcal{F}$ except that the equation for X is replaced by the equation $X = x$. The model $T_{\vec{X}=\vec{x}}$ describes a possible *counterfactual* situation; that is, even though, under normal circumstances, setting the exogenous variables to $\vec{u}$ results in the variables $\vec{X}$ having value $\vec{x}' \neq \vec{x}$, this submodel describes what happens if they are set to $\vec{x}$ due to some "external action," the cause of which is not modeled explicitly.

In general, given a *context* $\vec{u}$, that is, a setting for the exogenous variables, there may not be a unique vector of values that simultaneously satisfies all the equations in $T_{\vec{X}=\vec{x}}$; indeed, there may not be a solution at all. One special case where the equations in a causal model T are guaranteed to have a unique solution is when there is a total ordering $\prec_T$ of the variables in $\mathcal{V}$ such that if $X \prec_T Y$, then F_X is independent of the value of Y; that is, $F_X(\ldots, y, \ldots) = F_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{R}(Y)$. In this case, T is said to be *recursive* or *acyclic*. Intuitively, if T is recursive, then there is no feedback; if $X \prec_T Y$, then the value of X may affect the value of Y, but the value of Y has no effect on the value of X. It should be clear that if T is a recursive model, then, given a context $\vec{u}$, there is always a unique solution to the equations in $T_{\vec{X}=\vec{x}}$, for all $\vec{X}$ and $\vec{x}$, (We simply solve for the variables in the order given by $\prec_T$.)

Following Halpern (2000), I consider three successively more general classes of causal models for a given signature $\mathcal{S}$ (with a focus on the first two):

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

1. $\mathcal{T}_{\mathrm{rec}}(\mathcal{S})$: the class of recursive causal models over signature $\mathcal{S}$;
2. $\mathcal{T}_{\mathrm{uniq}}(\mathcal{S})$: the class of causal models T over $\mathcal{S}$ where, for all $\vec{X} \subseteq \mathcal{V}, \vec{x}$, the equations in $T_{\vec{X}=\vec{x}}$ have a unique solution for all contexts $\vec{u}$;
3. $\mathcal{T}(\mathcal{S})$: the class of all causal models over $\mathcal{S}$.

I often omit the signature $\mathcal{S}$ when it is clear from context or irrelevant, but the reader should bear in mind its important role.

Syntax and semantics. In Halpern (2000), a number of languages for reasoning about causality are considered. The choice of language is significant. As Galles and Pearl already point out, we cannot in any obvious way give a meaning in causal models to counterfactual implications where there is a disjunction on the left-hand side of the implication, that is formulas of the form $(A \vee A') \succeq B$. Thus, our results effectively consider a language with no disjunction on the left-hand side of $\succeq$. As mentioned in the introduction, one of the results of this paper have the form "for every recursive causal model, there is a recursive causal structure that satisfies the same formulas," and vice versa. For this result to have any bite, we must choose a reasonably rich language.

In Halpern (2000), I considered a number of languages appropriate for reasoning about causality in causal models. I briefly review two of them here. The languages are parameterized by the signature $\mathcal{S}$. A *basic causal formula* is one of the form $[Y_1 = y_1, \ldots, Y_k = y_k]\varphi$, where φ is a Boolean combination of formulas of the form $X = x$, $Y_1, \ldots, Y_k, X$ are variables in $\mathcal{V}$, $Y_1, \ldots, Y_k$ are distinct, and $x \in \mathcal{R}(X)$. I typically abbreviate such a formula as $[\vec{Y} = \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. A *causal formula* is a Boolean combination of basic causal formulas. Let $\mathcal{L}^+(\mathcal{S})$ consist of all causal formulas over the signature $\mathcal{S}$. (Again, I often omit the signature $\mathcal{S}$ if it is clear from context or not relevant.)

Roughly speaking, we can think of $\mathcal{L}^+(\mathcal{S})$ as the language that results by starting with primitive propositions of the form $X = x$, where X is a random variable in $\mathcal{V}$ and $x \in \mathcal{R}(X)$, and closing under modal operators of the form $[\vec{Y} = \vec{y}]$. The restriction to primitive propositions of this form is not a major one. Given a propositional language with primitive propositions $P_1, \ldots, P_n$, we can define binary random variables $X_1, \ldots, X_n$ (i.e., variables whose range is $\{0, 1\}$) and identify $X_i = 1$ with "P_i is true." That is, as long as we can we define structural equations that characterize how a change in one primitive propositions affects the other primitive propositions, taking the primitive propositions to have the form $X = x$ is not a major restriction.

The formula $[\vec{Y} = \vec{y}]X = x$ can be interpreted as "in all possible solutions to the structural equations obtained after setting Y_i to y_i, $i = 1, \ldots, k$, and the exogenous variables to $\vec{u}$, random variable X has value x." This formula is true in a causal model T in a context $\vec{u}$ if in all solutions to the equations in $T_{\vec{Y}=\vec{y}}$ in context $\vec{u}$, the random variable X has value x. Note that this formula is trivially true if there are no solutions to the structural equations.

A formula in $\mathcal{L}^+(\mathcal{S})$ is true or false in a causal model in $\mathcal{T}(\mathcal{S})$, given a context $\vec{u}$. As usual, we write $(T, \vec{u}) \models \varphi$ if the causal formula φ is true in causal model T given context $\vec{u}$.[3] For a basic causal formula $[\vec{Y} = \vec{y}]\varphi$, we define $(T, \vec{u}) \models [\vec{Y} = \vec{y}]\varphi$ if φ holds in all

[3] In Halpern (2000), following Galles & Pearl (1998), the context was included in the formula. The definition of $(T, \vec{u}) \models X = x$ given here is intended to be equivalent to that of $T \models X(\vec{u}) = x$. The advantage of having $\vec{u}$ on the left-hand side of $\models$ (which is also the formalism used in

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

solutions to the equations $\mathcal{F}^{\vec{Y}=\vec{y}}$ with the values of the variables in $\mathcal{U}$ set to $\vec{u}$. Thus, for example, if φ has the form $X_1 = x_1 \vee X_2 = x_2$, then $(T, \vec{u}) \models \varphi$ iff every vector of values for the endogenous variables that simultaneously satisfies all the equations in $\mathcal{F}$ has either $X_1 = x_1$ or $X_2 = x_2$.[4] We define the truth value of arbitrary causal formulas, which are just Boolean combinations of basic causal formulas, in the obvious way:

- $(T, \vec{u}) \models \varphi_1 \wedge \varphi_2$ if $(T, \vec{u}) \models \varphi_1$ and $(T, \vec{u}) \models \varphi_2$
- $(T, \vec{u}) \models \neg\varphi$ if $(T, \vec{u}) \not\models \varphi$.

As usual, a formula φ is said to be *valid* with respect to a class $\mathcal{T}'$ of causal models if $(T, \vec{u}) \models \varphi$ for all $T \in \mathcal{T}'$ and contexts $\vec{u}$ in T.

$\mathcal{L}^+$ is the most general language that I consider in Halpern (2000). To compare my results to those of GP, who use a more restricted language, I also consider some restrictions of $\mathcal{L}^+$. Specifically, let $\mathcal{L}_{\text{uniq}}$ be the sublanguage of $\mathcal{L}^+$ that consists of Boolean combinations of formulas of the form $[\vec{Y} = \vec{y}](X = x)$. Thus, the difference between $\mathcal{L}_{\text{uniq}}$ and $\mathcal{L}^+$ is that in $\mathcal{L}_{\text{uniq}}$, only $X = x$ is allowed after $[\vec{Y} = \vec{y}]$, while in $\mathcal{L}^+$, arbitrary Boolean combinations of formulas of the form $X = x$ are allowed. As the following lemma, proved in Halpern (2000), shows, for reasoning about causality in $\mathcal{T}_{\text{uniq}}$, the language $\mathcal{L}_{\text{uniq}}$ is adequate, since it is equivalent in expressive power to $\mathcal{L}^+$.

LEMMA 2.1. [Halpern, 2000] *In $\mathcal{T}_{\text{uniq}}$ and $\mathcal{T}_{\text{rec}}$, the language $\mathcal{L}^+$ and $\mathcal{L}_{\text{uniq}}$ are expressively equivalent; for every formula $\varphi \in \mathcal{L}^+$, we can effectively find a formula $\varphi' \in \mathcal{L}_{\text{uniq}}$ such that $\mathcal{T}_{\text{uniq}} \models \varphi \Leftrightarrow \varphi'$.*

The equivalence described in Lemma 2.1 no longer holds when reasoning about causality in the more general class $\mathcal{T}$ of structures. However, since I focus in this paper on $\mathcal{T}_{\text{rec}}$ and $\mathcal{T}_{\text{uniq}}$, in the rest of the paper I consider the language $\mathcal{L}_{\text{uniq}}$; this suffices to make my points.

Axiomatizations. I briefly recall some standard definitions from logic. An *axiom system* AX consists of a collection of *axioms* and *inference rules*. An axiom is a formula (in some predetermined language $\mathcal{L}$), and an inference rule has the form "from $\varphi_1, \ldots, \varphi_k$ infer ψ," where $\varphi_1, \ldots, \varphi_k, \psi$ are formulas in $\mathcal{L}$. A *proof* in AX consists of a sequence of formulas in $\mathcal{L}$, each of which is either an axiom in AX or follows by an application of an inference rule. A proof is said to be a *proof of the formula φ* if the last formula in the

Halpern & Pearl, 2005) is that it enforces the intuition that the context consists of background information that is typically suppressed. It does lead to loss in expressive power, since it is not possible to consider formulas of the form $X_1(\vec{u}_1) = x_1 \wedge X_2(\vec{u}_2) = x_2$, where $\vec{u}_1 \neq \vec{u}_2$. But this turns out to be an advantage—see the next footnote.

[4] In Halpern (2000), truth was defined for formulas of the form $[\vec{Y} = \vec{y}](X = x)$, and extended in "the obvious way" to Boolean combinations; that is, $\varphi_1 \vee \varphi_2$ was taken to be true if either φ_1 was true or φ_2 was true. This is equivalent to the approach described above as long as all equations have a unique solution, that is, in $\mathcal{T}_{\text{uniq}}$. However, the two approaches are not equivalent if equations can have more than one solution. The approach suggested here is what is required to make the axioms given in Halpern (2000) sound in the general case. We do not want to say that $X_1 = x_1 \vee X_2 = x_2$ is true if $X_1 = x_1$ is true in all solutions to the equations or $X_2 = x_2$ is true in all solutions to the equations; rather, we want either $X_1 = x_1$ or $X_2 = x_2$ to be true in all solutions. With the approach given above, it is not clear how to give semantics to formulas such as $X_1(\vec{u}_1) = x_1 \vee X_2(\vec{u}_2) = x_2$ if $\vec{u}_1 \neq \vec{u}_2$, which is perhaps another good reason for not allowing such formulas.

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

proof is φ. We say φ is *provable in AX*, and write $\mathrm{AX} \vdash \varphi$, if there is a proof of φ in AX; similarly, we say that φ is *consistent with AX* if $\neg\varphi$ is not provable in AX.

An axiom system AX is said to be *sound* for a language $\mathcal{L}$ with respect to a class $\mathcal{T}'$ of causal models if every formula in $\mathcal{L}$ provable in AX is valid with respect to $\mathcal{T}'$. AX is *complete* for $\mathcal{L}$ with respect to $\mathcal{T}'$ if every formula in $\mathcal{L}$ that is valid with respect to $\mathcal{T}'$ is provable in AX.

Consider the following axioms, taken from Halpern (2000), modified slightly for the language used here:

C0. All instances of propositional tautologies in the language $\mathcal{L}_{\mathrm{uniq}}$.

C1. $[\vec{Y} = \vec{y}](X = x) \Rightarrow \neg[\vec{Y} = \vec{y}](X = x')$, if $x, x' \in \mathcal{R}(X), x \neq x'$. (Equality)

C2. $\vee_{x \in \mathcal{R}(X)}[\vec{Y} = \vec{y}](X = x)$. (Definiteness)

C3. $([\vec{X} = \vec{x}](W = w) \wedge ([\vec{X} = \vec{x}](Y = y)) \Rightarrow [\vec{X} = \vec{x}; W = w](Y = y))$.[5] (Composition)

C4. $[X = x; \vec{W} = \vec{w}](X = x)$. (Effectiveness)

C5. $([\vec{X} = \vec{x}; W = w](Y = y) \wedge [\vec{X} = \vec{x}; Y = y](W = w)) \Rightarrow [\vec{X} = \vec{x}](Y = y)$, if $Y \neq W$.[6] (Reversibility)

The key axioms C3–C5 were introduced (and named) by Galles & Pearl (1998). Perhaps most relevant to this paper is the reversibility axiom, C5. It says that if setting $\vec{X}$ to $\vec{x}$ and W to w results in Y having value y and setting $\vec{X}$ to $\vec{x}$ and Y to y results in W having value w, then Y must already have value y when we set $\vec{X}$ to x (and W must already have value w).

Let $\mathrm{AX}_{\mathrm{uniq}}(\mathcal{S})$ consist of C0–C5 and the rule of inference *modus ponens* (from φ and $\varphi \Rightarrow \psi$ infer ψ).

THEOREM 2.2. [Halpern, 2000] *$AX_{\mathrm{uniq}}(\mathcal{S})$ is a sound and complete axiomatization for $\mathcal{L}_{\mathrm{uniq}}(\mathcal{S})$ with respect to $\mathcal{T}_{\mathrm{uniq}}(\mathcal{S})$.*

Using Lemma 2.1, it is possible to get a complete axiomatization for $\mathcal{L}^+$ with respect to $\mathcal{T}_{\mathrm{uniq}}(\mathcal{S})$, simply by adding axioms for converting a formula in $\mathcal{L}^+$ to an equivalent formula in $\mathcal{L}_{\mathrm{uniq}}$. A complete axiomatization for $\mathcal{L}_{\mathrm{uniq}}$ with respect to $\mathcal{T}_{\mathrm{rec}}(\mathcal{S})$ is also given in Halpern (2000); it requires adding a somewhat complicated axiom called C6 to $\mathrm{AX}_{\mathrm{uniq}}$ that captures acyclicity. The details are not relevant for our purposes, so I do not discuss C6 further.

Finally, a complete axiomatization for $\mathcal{L}^+$ with respect to $\mathcal{T}(\mathcal{S})$, the class of all causal models, is given. The axioms are similar in spirit to those in $\mathrm{AX}_{\mathrm{uniq}}$. In particular, there is the following analogue to reversibility (where $\langle \vec{X} = \vec{x} \rangle \varphi$ is an abbreviation for $\neg[\vec{X} = \vec{x}]\neg\varphi$):

$$(\langle \vec{X} = \vec{x}; Y = y \rangle (W = w \wedge \vec{Z} = \vec{z}) \wedge \langle \vec{X} = \vec{x}; W = w \rangle (Y = y \wedge \vec{Z} = \vec{z})) \Rightarrow \langle \vec{X} = \vec{x} \rangle (W = w \wedge Y = y \wedge \vec{Z} = \vec{z})), \quad \text{where } \vec{Z} = \mathcal{V} - (\vec{X} \cup \{W, Y\}).$$

[5] Galles and Pearl use a stronger version of C3: $([\vec{X} = \vec{x}](W = w) \Rightarrow ([\vec{X} = \vec{x}](Y = y)) \Leftrightarrow [\vec{X} = \vec{x}; W = w](Y = y))$. The stronger version follows from the weaker version in the presence of the other axioms. (This is actually shown in the proof of Proposition 3.3.)

[6] The assumption that $Y \neq W$ was not made by Galles & Pearl (1998) nor by Halpern (2000), but it is necessary. For example, if $W = Y$, then one instance of C5 would be $([\vec{X} = \vec{x}; Y = y](Y = y) \wedge [\vec{X} = \vec{x}; Y = y](Y = y)) \Rightarrow [\vec{X} = \vec{x}](Y = y)$. The antecedent is true by C4, while the conclusion is not true in general.

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

2.2. Possible-worlds models for counterfactuals. There have been a number of semantics for counterfactuals. I focus here on one due to Lewis (1973). Let Φ be a finite set of primitive propositions. A *counterfactual structure* M is a tuple (Ω, R, π), where Ω is a finite set of *possible worlds*,[7] π is an *interpretation* that maps each possible world to a truth assignment over Φ, and R is a ternary relation over Ω. Intuitively, $(w, u, v) \in R$ if u is as close/preferred/plausible as v when the real world is w. Let $u \preceq_w v$ be an abbreviation for $(w, u, v) \in R$, and define $\Omega_w = \{u : u \preceq_w v \in R \text{ for some } v \in \Omega\}$; thus, the worlds in Ω_w are those that are at least as plausible as some world in Ω according to $\preceq_w$. Define $v \prec_w v'$ if $v \preceq_w v'$ and $v' \not\preceq_w v$. We require that $w \in \Omega_w$, that $\preceq_w$ be reflexive and transitive on Ω_w, and that $w \prec_w u$ for all $u \neq w$. Thus, $\prec_w$ puts an ordering on worlds that can be viewed as characterizing "closeness to w," and w is the closest world to itself.

Let $\mathcal{L}^C(\Phi)$ be the language formed by starting with Φ and closing off under $\wedge$, $\neg$, and $\succeq$ (where $\succeq$ denotes counterfactual implication). The language $\mathcal{L}^C$ allows arbitrary nesting of counterfactual implications. By way of contrast, the language $\mathcal{L}^+$ and its sublanguages have only one level of nesting, if we think of $[\vec{X} = \vec{x}]\varphi$ as $(\vec{X} = \vec{x}) \succeq \varphi$. Let $\mathcal{L}^C_1$ be the sublanguage of $\mathcal{L}^C$ consisting of all formulas with no nested occurrence of $\succeq$ (including formulas with no occurrence of $\succeq$ at all).

We can give semantics to formulas in $\mathcal{L}^C$ (and hence $\mathcal{L}^C_1$) in a counterfactual structures $M = (\Omega, R, \pi)$ as follows. The first few clauses are standard:

- $(M, w) \models p$, when $p \in \Phi$, if $\pi(w)(p) = \mathbf{true}$.
- $(M, w) \models \varphi \wedge \psi$ if $(M, w) \models \varphi$ and $(M, w) \models \psi$.
- $(M, w) \models \neg\varphi$ if it is not the case that $(M, w) \models \varphi$.

To give semantics to $\varphi \succeq \psi$, assume inductively that we have already given semantics to φ at all worlds in M. Define $closest_M(w, \varphi) = \{v \in \Omega_w : (M, v) \models \varphi$ and there is no world v' such that $(M, v') \models \varphi$ and $v' \prec_w v\}$. Thus, $closest_M(w, \varphi)$ is the set of worlds closest to w where φ is true. (Notice that if there are no worlds where φ holds, then $closest_M(w, \varphi) = \emptyset$.)

- $(M, w) \models \varphi \succeq \psi$ if $(M, v) \models \psi$ for all $v \in closest_M(w, \varphi)$.

Axioms. There are a number of well-known sound and complete axiomatizations for counterfactual logic (see, e.g., Burgess, 1981; Bell, 1989; Chellas, 1980; Grahne, 1991; Katsuno & Satoh, 1991; Lewis, 1971, 1973, 1974). Here is one, based on Burgess's axiomatization, similar in spirit to the well-known KLM properties (Kraus *et al.*, 1990).

A0. All instances of propositional tautologies in the language $\mathcal{L}^C$.
A1. $\varphi \succeq \varphi$.
A2. $((\varphi \succeq \psi_1) \wedge (\varphi \succeq \psi_2)) \Rightarrow (\varphi \succeq (\psi_1 \wedge \psi_2))$.
A3. $((\varphi_1 \succeq \varphi_2) \wedge (\varphi_1 \succeq \psi)) \Rightarrow ((\varphi_1 \wedge \varphi_2) \succeq \psi)$.
A4. $((\varphi_1 \succeq \psi) \wedge (\varphi_2 \succeq \psi)) \Rightarrow ((\varphi_1 \vee \varphi_2) \succeq \psi)$.
A5. $\neg(\mathit{true} \succeq \mathit{false})$.
A6. $\varphi \Rightarrow (\psi \Leftrightarrow (\varphi \succeq \psi))$.

[7] Giving semantics to counterfactual formulas in structures with infinitely many worlds adds an extra level of complexity. As shown by Friedman & Halpern (1994), if a formula in the language I am about to introduce is satisfiable at all, it is satisfiable in a structure with finitely many worlds, so as far as validity goes, there is no loss of generality in restricting to finite structures.

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

There are three rules of inference: modus ponens, and the following two rules:

RA1. From $\varphi \Leftrightarrow \varphi'$ infer $(\varphi \succeq \psi) \Rightarrow (\varphi' \succeq \psi)$.
RA2. From $\psi \Rightarrow \psi'$ infer $(\varphi \succeq \psi) \Rightarrow (\varphi \succeq \psi')$.

A1–A4 and RA1–RA2 correspond to the KLM postulates REF (for reflexivity), AND, CM (cautious monotonicity), OR, LLE (left logical equivalence), and RW (right weakening), respectively. A5 captures the requirement that Ω_w is nonempty. Finally, A6 is the axiom that makes $\succeq$ a counterfactual operator, and not just a "normality" or "typicality" operator (so that "normally birds have wings" becomes *bird* $\succeq$ *wing*). Suppose that it is the case that if φ were true, then ψ would be true. Then if φ is actually true, we would expect ψ to be true. Moreover, if φ and ψ are both true, then it seems reasonable to assert that if φ were true, then ψ would be true. On the other hand, it is not in general the case that if φ and ψ are both true, then φ's are normally or typically ψ's.

In his semantics, Lewis allows there to be more than one world closest to a world w where φ is true (except in the special case that φ is true at w; in this case, w is the unique world closest to w satisfying φ). By way of contrast, Stalnaker (1968) essentially assumes that for each world w and formula φ, there is a unique world closest to w satisfying φ. The standard approach to getting uniqueness is to require that $\preceq_w$ be a strict total order (whose least element is w); that is, for all worlds $w' \neq w''$, either $w' \prec_w w''$ or $w'' \prec_w w'$. This assumption is captured by the following axiom:

A7. $(\varphi \succeq (\psi_1 \vee \psi_2)) \Rightarrow ((\varphi \succeq \psi_1) \vee (\varphi \succeq \psi_2))$.

Let AX be the axiom system consisting of axioms A0–A6 and rules of inference RA1, RA2, and modus ponens; let AX′ be AX together with the axiom A7. Let $\mathcal{M}(\Phi)$ be the collection of all counterfactual structures over the primitive propositions in Φ (i.e., structures where π interprets formulas in Φ); let $\mathcal{M}^+(\Phi)$ be the subset of $\mathcal{M}(\Phi)$ consisting of all counterfactual structures where $\preceq_w$ is a total strict order. As usual, I omit the Φ when it is clear from context or irrelevant.

THEOREM 2.3. [Burgess, 1981] *AX (resp., AX′) is a sound and complete axiomatization for the language* $\mathcal{L}_{\text{uniq}}$ *with respect to* $\mathcal{M}$ *(resp.,* $\mathcal{M}^+$*).*

§3. Relating causal models to counterfactual structures. As the title suggests, in this section I take a closer look at the relationship between causal models and counterfactual structures. On the surface, the two approaches are quite different. Consider the causal model for the forest fire example, discussed in Section 2.1. If we want to capture the forest fire using counterfactual structures, perhaps the most natural way to do it is to have worlds in the counterfactual structure that correspond to the eight possible settings of the three exogenous variables (*MD*, *L*, and *FF*). We can take primitive propositions that correspond to the settings of these variables as well; that is, the primitive propositions have the form $MD = i$, $L = i$, and $FF = i$, for $i = 0, 1$. The actual world w is the one where $MD = 1 \wedge L = 1 \wedge FF = 1$ holds. The closest world relation is described in the obvious way by the equations. For example, if we consider the conjunctive model, where $FF = \min(MD, L)$, so that both the match and the lightning are required to start the fire, then the closest world to w where $MD = 0$ is the one where $MD = 0 \wedge L = 1 \wedge FF = 0$ holds. On the other hand, in the disjunctive model, where $FF = \max(MD, L)$, the closest world to w where $MD = 0$ holds is the one where $MD = 0 \wedge L = 1 \wedge FF = 1$ holds. There is a sense in which the causal model and the corresponding counterfactual

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

structure(s) constructed this way satisfy the same formulas. In this section, I make this intuition precise. More generally, I show that to every causal model in $\mathcal{T}_{\text{rec}}$, there is a corresponding counterfactual structure that satisfies the same formulas; however, this is not the case for every causal model in $\mathcal{T}_{\text{uniq}}$.

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, consider the set $\Phi_{\mathcal{S}}$ of primitive propositions of the form $X = x$ for $X \in \mathcal{V}$ and $x \in \mathcal{R}(x)$. We restrict to counterfactual structures $M = (\Omega, R, \pi)$ for this set of primitive propositions, where π is such that, for each world $w \in \Omega$ and variable $X \in \mathcal{V}$, exactly one of the formulas $X = x$ is true. Call such structures *acceptable*. In an acceptable counterfactual structure, a world can be associated with an assignment of values to the random variables. An acceptable structure (Ω, R, π) is *full* if, for each assignment v of values to variables and all $w \in \Omega$, there is a world in Ω_w where v is the assignment. (I discuss the consequences of fullness shortly.) Let $\mathcal{M}_a(\Phi_{\mathcal{S}})$ consist of all acceptable counterfactual structures over $\mathcal{S}$; let $\mathcal{M}_f(\Phi_{\mathcal{S}})$ consist of all full acceptable counterfactual structures over $\mathcal{S}$; let $\mathcal{M}_a^+(\Phi_{\mathcal{S}}) = \mathcal{M}_a(\Phi_{\mathcal{S}}) \cap \mathcal{M}^+(\Phi_{\mathcal{S}})$; and let $\mathcal{M}_f^+(\Phi_{\mathcal{S}}) = \mathcal{M}_f(\Phi_{\mathcal{S}}) \cap \mathcal{M}^+(\Phi_{\mathcal{S}})$.

As before, I identify $[\vec{Y} = \vec{y}](X = x) \in \mathcal{L}_{\text{uniq}}(\mathcal{S})$ with the formula $\vec{Y} = \vec{y} \succeq (X = x) \in \mathcal{L}_1^C(\Phi_{\mathcal{S}})$. I abuse notation and use $\mathcal{L}_{\text{uniq}}(\mathcal{S})$ to denote the sublanguage of $\mathcal{L}_1^C(\Phi_{\mathcal{S}})$ that arises via this identification. The following is easy to show.

PROPOSITION 3.1. *C3 and C4 are valid in $\mathcal{M}_a(\Phi_{\mathcal{S}})$; C2 is valid in $\mathcal{M}_a^+(\Phi_{\mathcal{S}})$; C1 is valid in $\mathcal{M}_f(\varphi_{\mathcal{S}})$.*

Notice that, in $\mathcal{M}_a(\Phi_{\mathcal{S}})$ (and hence all of its subsets), the following two formula schemes are valid:

V1. $\vee_{x \in \mathcal{R}(X)} X = x$
V2. If $x \neq x'$, then $X = x \Rightarrow X \neq x'$.

In $\mathcal{M}_f(\Phi_{\mathcal{S}})$, the following axiom is valid:

V3. $\neg[\vec{X} = \vec{x}]\mathit{false}$.

V3 is valid in a structure where, for all worlds w, there is a closest world to w such that $\vec{X} = \vec{x}$, that is, $\mathit{closest}_M(w, \vec{X} = \vec{x}) \neq \emptyset$, which is exactly what fullness ensures.[8]

With this background, I can give an "axiomatic" proof of Proposition 3.1:

- C1 follows easily from A0, A2, A5, and V2.
- C3 is a special case of A3.
- C4 follows easily from A1 and RA2.
- C2 is not provable in AX (and is not valid in $\mathcal{M}_a(\Phi_{\mathcal{S}})$), but it is provable in AX′ together with the axiom V1. Indeed, it follows easily from A1, A7, RA2, and V1 (since $\varphi \Rightarrow \vee_{x \in \mathcal{R}(X)} X = x$ is valid in $\mathcal{M}(\Phi_{\mathcal{S}})$). The fact that C2 requires A7 is not surprising. C2 is essentially expressing the uniqueness of solutions, which is being captured by the assumption that there is a unique closest world where the antecedent is true, an assumption captured by A7.

Reversibility (C5) is conspicuously absent from this list. Indeed, as the following example shows, C5 is not sound in counterfactual structures.

[8] Galles and Pearl did not discuss the axioms V1–V3, but it is clear that they are implicitly assuming that they hold.

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

EXAMPLE 3.2. *Suppose that* $\mathcal{V} = \{X_1, X_2, X_3\}$, *and* $\mathcal{R}(X_1) = \mathcal{R}(X_2) = \mathcal{R}(X_3) = \{0, 1\}$. *Consider a structure* $(\Omega, R, \pi) \in \mathcal{M}_f^+$ *where there is exactly one world in* Ω *corresponding to each of the 8 possible assignments of values to the variables. Thus, we can identify a world* w *in* Ω *with a triple* (b_1, b_2, b_3), *where, in* w, $X_i = b_i$. *All that matters about* R *is that* $(0, 0, 0) \prec_{(0,0,0)} (1, 0, 0) \prec_{(0,0,0)} (1, 1, 1)$, *with the other 5 worlds being further from* $(0, 0, 0)$ *than* $(1, 1, 1)$. *Then it is immediate that*

$$(M, (0, 0, 0)) \models [X_1 = 1; X_2 = 1](X_3 = 1) \wedge [X_1 = 1; X_3 = 1](X_2 = 1) \wedge [X_1 = 1](X_2 = 0).$$

Thus, C5 is violated.

For the remainder of this section, fix $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$. Say that a counterfactual structure $M = (\Omega, R, \pi)$ in $\mathcal{M}_f^+(\Phi_{\mathcal{S}})$ is *recursive* if there is a total ordering $\prec_M$ of the variables in $\mathcal{V}$ such that if $W \prec_M Y$, then for all $\vec{X} \subseteq \mathcal{V} - \{Y, W\}$, for each world $w \in \Omega$, in the closest world to w where $\vec{X} = \vec{x}$ and $Y = y$, the value of W is the same as in the closest world to w where $\vec{X} = \vec{x}$. Intuitively, setting Y to y has no further effect on the value of W once $\vec{X}$ is set to $\vec{x}$. Let $\mathcal{M}_{\mathrm{rec}}$ consist of the recursive structures in $\mathcal{M}_f^+$. It is easy to see that the structure M considered in Example 3.2 is not in $\mathcal{M}_{\mathrm{rec}}$. For if it were, then we would have either $X_2 \prec_M X_3$ or $X_3 \prec_M X_2$. If $X_3 \prec_M X_2$, then the value of X_3 would have to be the same in the closest world to $(0, 0, 0)$ where $X_1 = 1$ and $X_2 = 1$ as in the closest world to $(0, 0, 0)$ where $X_1 = 1$. But it is not. The same problem occurs if $X_2 \prec_M X_3$.

PROPOSITION 3.3. *C5 is valid in* $\mathcal{M}_{\mathrm{rec}}$.

Proof. Suppose that $M = (\Omega, R, \pi) \in \mathcal{M}_{\mathrm{rec}}$, $w \in \Omega$, and the antecedent of C5 holds at (M, w). If $Y \prec_M W$, then it is immediate from the fact that $(M, w) \models [\vec{X} = \vec{x}; Y = y](W = w)$ that we also have $(M, w) \models [\vec{X} = \vec{x}](W = w)$. Now suppose that $W \prec_M Y$. Then from $(M, w) \models [\vec{X} = \vec{x}; W = w](Y = y)$ it follows that $(M, w) \models [\vec{X} = \vec{x}](Y = y)$. Suppose, by way of contradiction, that $(M, w) \models [\vec{X} = \vec{x}](W = w')$ for some $w' \neq w$. By C3 (which, by Proposition 3.1, is sound in M), it follows that $(M, w) \models [\vec{X} = \vec{x}; Y = y](W = w')$. This, combined with the fact that $(M, w) \models [\vec{X} = \vec{x}; Y = y](W = w)$, contradicts C1 (which, by Proposition 3.1, is also sound in M). (Note that this argument actually shows that the stronger version of C3 used by Galles and Pearl, discussed in Footnote 4, follows from the weaker version used here.) □

Not surprisingly, the proof of Proposition 3.3 is essentially identical to the argument given by Galles and Pearl that reversibility holds in recursive structures. Indeed, it is not hard to show that the axiom C6 that characterizes recursive structures is valid in $\mathcal{M}_{\mathrm{rec}}$, and C5 follows from the other axioms in the presence of C6.

Even more can be shown. In a precise sense, every causal model in $\mathcal{T}_{\mathrm{rec}}(\mathcal{S})$ can be identified with a counterfactual structure in $\mathcal{M}_{\mathrm{rec}}(\mathcal{S})$ where the same formulas are true. It follows that recursive counterfactual structures are at least as general as recursive causal models. The converse is also true. I now make these claims precise.

Given a causal model $T = (\mathcal{S}, \mathcal{F}) \in \mathcal{T}_{\mathrm{rec}}$, we construct a model $M_T = (\Omega, R, \pi)$ as follows. Let Ω consist of all the assignments of values to the variables in $\mathcal{U} \cup \mathcal{V}$. The interpretation π is defined in the obvious way: $X = x$ is true in w if w assigns X value x. For each context $\vec{u}$, let $\vec{v}_{\vec{u}}$ be the assignment to the variables in $\mathcal{V}$ that is forced by the equations. (Since $T \in \mathcal{T}_{\mathrm{rec}}$, $\vec{v}_{\vec{u}}$ is uniquely determined.) More generally, for each assignment $\vec{X} = \vec{x}$, let $\vec{v}_{\vec{u}, \vec{X} = \vec{x}}$ be the assignment to the variables in V determined by the

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

equations $\mathcal{F}^{\vec{X}=\vec{x}}$ in the context $\vec{u}$. Let $w_{\vec{u}} = (\vec{u}, \vec{v}_{\vec{u}})$ and let $w_{\vec{u},\vec{X}=\vec{x}} = (\vec{u}, \vec{v}_{\vec{u},\vec{X}=\vec{x}})$. Finally, let R be such that the closest world to $w_{\vec{u}}$ where $\vec{X} = \vec{x}$ is $w_{u,\vec{X}=\vec{x}}$, and for all assignments $\vec{u}$ and $\vec{v}$ to the exogenous and endogenous variables, respectively, $\Omega_{\vec{u},\vec{v}}$ consists of all worlds $(\vec{u}, \vec{v}')$ such that $\vec{v}'$ is an arbitrary assignment to the endogenous variables. This does not uniquely determine R. Indeed, it places no constraints on $\preceq_w$ if w is not of the form $w_{\vec{u}}$ and does not completely determine R even if w does have the form $w_{\vec{u}}$. It is easy to define a relation R that satisfies these constraints such that (1) $\prec_{w_{\vec{u}}}$ is a strict total order on $\Omega_{w_{\vec{u}}}$ and (2) for w not of the form $w_{\vec{u}}$, $\prec_w$ is a strict total order on Ω_w that satisfies the recursiveness constraint.

THEOREM 3.4. *$M_T \in \mathcal{M}_{\mathrm{rec}}$ and, for all formulas $\varphi \in \mathcal{L}_{\mathrm{uniq}}(\mathcal{S})$, we have $(T, \vec{u}) \models \varphi$ iff $(M_T, w_{\vec{u}}) \models \varphi$.*

Proof. It is easy to see that we can take $\prec_{M_T} = \prec_T$, so $M_T \in \mathcal{M}_{\mathrm{rec}}$. Using the definition of R, it is easy to show by induction on the structure of formulas that $(T, \vec{u}) \models \varphi$ iff $(M_T, w_{\vec{u}}) \models \varphi$. I leave the details to the reader. □

Theorem 3.4 shows that we can embed $\mathcal{T}_{\mathrm{rec}}$ in $\mathcal{M}_{\mathrm{rec}}$. I next give an embedding of $\mathcal{M}_{\mathrm{rec}}$ in $\mathcal{T}_{\mathrm{rec}}$. Now the causal model in $\mathcal{T}_{\mathrm{rec}}$ depends both on the counterfactual structure and a world in that structure. (I discuss why this is so after Theorem 3.5.) Suppose that $M = (\Omega, R, \pi) \in \mathcal{M}_{\mathrm{rec}}(\mathcal{S})$ and $w \in \Omega$. Suppose that $\mathcal{V} = \{X_1, \ldots, X_n\}$ and, without loss of generality, that $X_i \prec_M X_j$ iff $i < j$. Consider the causal model $T_{M,w} = (\mathcal{S}, \mathcal{F})$, where F_X is defined by induction on the $\prec_M$-ordering. That is, we first define F_{X_1}, since X_1 is the $\prec$-minimal variable, then define F_{X_2}, and so on. Suppose that $(M, w) \models X_1 = x_1$. Then define F_{X_1} so that, for all contexts $\vec{u}$ and assignments $\vec{v}$ to the variables in $\mathcal{V} - \{X_1\}$, $F_{X_1}(\vec{u}, \vec{v}) = x_1$. Suppose that we have defined F_{X_i} for $i \leq k$. Define $F_{X_{k+1}}$ so that, for all contexts $\vec{u}$ and all assignments $\vec{v}$ to the variables in $\mathcal{V} - \{X_{k+1}\}$, we have $F_{X_{k+1}}(\vec{u}, \vec{v}) = x$ iff $(M, w) \models [X_1 = v_1; \ldots, X_k = v_k](X_{k+1} = x)$.

THEOREM 3.5. *$T_{M,w} \in \mathcal{T}_{\mathrm{rec}}$ and, for all formulas $\varphi \in \mathcal{L}_{\mathrm{uniq}}(\mathcal{S})$ and all contexts $\vec{u}$, we have $(M, w) \models \varphi$ iff $(T_{M,w}, \vec{u}) \models \varphi$.*

Proof. It is easy to see that $T_{M,w} \in \mathcal{T}_{\mathrm{rec}}$; the definition of $\mathcal{F}$ guarantees that $\prec_{T_{M,w}} = \prec_M$. The definition of F_X is independent of the context; it easily follows that for all contexts $\vec{u}$ and $\vec{u}'$, we have $(T_{M,w}, \vec{u}) \models \varphi$ iff $(T_{M,w}, \vec{u}') \models \varphi$. Again, an easy induction on the structure of φ shows that, for all contexts $\vec{u}$, $(M, w) \models \varphi$ iff $(T_{M,w}, \vec{u}) \models \varphi$. □

It is easy to modify the construction slightly so that each context $\vec{u}$ corresponds to a different world $w \in \Omega$. Thus, if the number of contexts is at least $|\Omega|$, then we can get a closer analogue to Theorem 3.4, where we can associate with each world $w \in \Omega$ a context $\vec{u}_w$ and show that $(M, w) \models \varphi$ iff $(T, \vec{u}_w) \models \varphi$. However, Theorem 3.5 suffices for the following corollary.

COROLLARY 3.6. *The same formulas in $\mathcal{L}_{\mathrm{uniq}}(\Phi_{\mathcal{S}})$ are valid in both $\mathcal{T}_{\mathrm{rec}}(\mathcal{S})$ and $\mathcal{M}_{\mathrm{rec}}(\mathcal{S})$.*

Proof. If φ is not valid in $\mathcal{T}_{\mathrm{rec}}$, then there is some causal model $T \in \mathcal{T}_{\mathrm{rec}}$ and context $\vec{u}$ such that $(T, \vec{u}) \models \neg\varphi$. By Theorem 3.4, $(M_T, w_{\vec{u}}) \models \neg\varphi$, so φ is not valid in $\mathcal{M}_{\mathrm{rec}}$. For the converse, if φ is not valid in $\mathcal{M}_{\mathrm{rec}}$, then there is some counterfactual structure $M \in \mathcal{M}_{\mathrm{rec}}$ and world w such that $(M, w) \models \neg\varphi$. By Theorem 3.5, $(T_{M,w}, \vec{u}) \models \neg\varphi$. Thus, the same formulas are satisfiable in both $\mathcal{T}_{\mathrm{rec}}$ and $\mathcal{M}_{\mathrm{rec}}$, and hence the same formulas are valid. □

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

What happens if we consider $\mathcal{T}_{\text{uniq}}$ rather than $\mathcal{T}_{\text{rec}}$? Now causal models and counterfactual structures are incomparable. Consider the formula

$$\begin{aligned}\varphi =_{\text{def}} \quad & [X_1 = 1](X_2 = 1 \wedge X_3 = 0) \wedge \\ & [X_2 = 1](X_3 = 1 \wedge X_1 = 0) \wedge \\ & [X_3 = 1](X_1 = 1 \wedge X_2 = 0).\end{aligned}$$

THEOREM 3.7. *$\neg\varphi$ is valid in $\mathcal{T}_{\text{rec}}$ and $\mathcal{M}_f$ (and hence $\mathcal{M}_f^+$), but φ is satisfiable in $\mathcal{T}_{\text{uniq}}$.*

Proof. The validity of $\neg\varphi$ in $\mathcal{T}_{\text{rec}}$ follows from its validity in $\mathcal{M}_f$, in light of Corollary 3.6 and the fact that $\mathcal{M}_{\text{rec}} \subseteq \mathcal{M}_f$. Nevertheless, I prove the validity of $\neg\varphi$ in $\mathcal{T}_{\text{rec}}$ first, since the proof is short and gives some insight. Consider a causal model $T \in \mathcal{T}_{\text{rec}}$, and suppose, by way of contradiction, that $(M, w) \models \varphi$. Let $\prec_T$ be the total ordering on the variables in $\mathcal{V}$ in T. One of X_1, X_2, and X_3 must be minimal with respect to $\prec_T$. Suppose it is X_1 and that $(T, \vec{u}) \models X_1 = i$. Then we must have $(T, \vec{u}) \models [X_2 = 1](X_1 = i)$ and $(T, \vec{u}) \models [X_3 = 1](X_1 = i)$. But since $(T, \vec{u}) \models \varphi$, it follows that $(T, \vec{u}) \models [X_2 = 1](X_1 = 0)$ and $(T, \vec{u}) \models [X_3 = 1](X_1 = 1)$. Thus, X_1 cannot be minimal with respect to $\prec_T$. An analogous argument shows that X_2 and X_3 also cannot be minimal with respect to $\prec_T$. Thus, we have a contradiction.

I next show that $\neg\varphi$ is valid in $\mathcal{M}_f$. Suppose by way of contradiction that $M \in \mathcal{M}_f$ and $(M, w) \models \varphi$. Consider a world w' closest to w that satisfies $X_1 = 1 \vee X_2 = 1 \vee X_3 = 1$. (Since $M \in \mathcal{M}_f$, there is guaranteed to be such a world.) Suppose that $(M, w') \models X_1 = 1$. Note that w' must be one of the worlds closest to w that satisfies $X_1 = 1$. Since $(M, w) \models [X_1 = 1](X_2 = 1)$, we must have $(M, w') \models X_2 = 1$. Thus, w' must also be one of the worlds closest to w that satisfies $X_2 = 1$. Since $(M, w) \models [X_2 = 1](X_3 = 1)$, we must have that $(M, w') \models X_3 = 1$. On the other hand, since $(M, w) \models [X_1 = 1](X_3 = 0)$, we must have have that $(M, w') \models X_3 = 0$. This gives a contradiction. A similar contradiction arises if $(M, w') \models X_2 = 1$ or if $(M, w') \models X_3 = 1$. Since $(M, w') \models X_1 = 1 \vee X_2 = 1 \vee X_3 = 1$ by construction, this gives a contradiction to the assumption that $(M, w) \models \varphi$.

Finally, I must show that there is a causal model $T \in \mathcal{T}_{\text{uniq}}$ and a context $\vec{u}$ such that $(T, \vec{u}) \models \varphi$. Let $\mathcal{V} = \{X_1, X_2, X_3\}$ and let $\mathcal{U} = \{U\}$. Take $\mathcal{R}(X_1) = \mathcal{R}(X_2) = \mathcal{R}(X_3) = \{0, 1\}$ and $\mathcal{R}(U) = \{0\}$. In defining $\mathcal{F}$, I write F_i instead of F_{X_i}, and omit the U argument (since it is always 0). Thus, $F_1(0, 0) = 0$ means that when $X_2 = X_3 = 0$, then $X_1 = 0$. Define

- $F_1(0, 0) = 0$; $F_1(0, 1) = 1$; $F_1(1, 0) = 0$; $F_1(1, 1) = 0$;
- $F_2(0, 0) = 0$; $F_2(0, 1) = 0$; $F_2(1, 0) = 1$; $F_2(1, 1) = 0$;
- $F_3(0, 0) = 0$; $F_3(0, 1) = 1$; $F_3(1, 0) = 0$; $F_3(1, 1) = 0$.

I must now verify that $T = (\mathcal{S}, \mathcal{F}) \in \mathcal{T}_{\text{uniq}}$, and that $(T, 0) \models \varphi$. This is straightforward, although tedious. First observe that $(0, 0, 0)$ is the only solution to all the equations in the basic causal model. It is easy to see that $(0, 0, 0)$ is a solution. To see that it is the only solution, observe that $(0, 0, 1)$ cannot be a solution because $F_3(0, 0) = 0$; similarly, $(0, 1, 0)$ and $(1, 0, 0)$ cannot be solutions; $(0, 1, 1)$ cannot be a solution because $F_2(0, 1) = 0$; $(1, 1, 0)$ cannot be a solution because $F_1(1, 0) = 0$; $(1, 0, 1)$ cannot be a solution because $F_3(1, 0) = 0$; finally, $(1, 1, 1)$ cannot be a solution because $F_1(1, 1) = 0$.

It is clear that there must be a unique solution if we set two of the three variables (forced by the equation for the third variable); for example, if $X_1 = 1$ and $X_2 = 0$, then the unique solution is $(1, 0, 0)$. If X_1, X_2, or X_3 is set to 0, then one solution is $(0, 0, 0)$. The solution is unique for the same reasons that $(0, 0, 0)$ was the unique solution to the original collection

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

of equations. Finally, if $X_1 = 1$, then $(1, 1, 0)$ is a solution; if $X_2 = 1$, then $(0, 1, 1)$ is a solution; and if $X_3 = 1$, then $(1, 0, 1)$ is a solution. We must show that there are no other solutions in all three cases. If $X_1 = 1$, (1,0,0) is not a solution since $F_2(1, 0) = 1$; $(1, 0, 1)$ is not a solution because $F_3(1, 0) = 0$; and $(1, 1, 1)$ is not a solution because $F_3(1, 1) = 0$. If $X_2 = 1$, $(0, 1, 0)$ is not a solution because $F_3(0, 1) = 1$; $(1, 1, 0)$ is not a solution because $F_1(1, 0) = 0$; $(1, 1, 1)$ is not a solution because $F_1(1, 1) = 0$. Finally, if $X_3 = 1$, $(0, 0, 1)$ is not a solution because $F_1(0, 1) = 1$; $(0, 1, 1)$ is not a solution because $F_2(0, 1) = 0$; and $(1, 1, 1)$ is not a solution because $F_1(1, 1) = 0$. Thus, $T \in \mathcal{T}_{\text{uniq}}$.

It is now straightforward to check that $(T, u) \models \varphi$. □

In Section 2.1, I argued that the choice of language was significant. All the results of this section are stated for the language $\mathcal{L}_{\text{uniq}}$. In light of Lemma 2.1, Theorems 3.4 and 3.5 apply without change to the language $\mathcal{L}^+$. Moreover, since $\mathcal{T}_{\text{uniq}} \subseteq \mathcal{T}$ and $\mathcal{L}_{\text{uniq}} \subseteq \mathcal{L}^+$, the formula $\neg\varphi$ from Theorem 3.7 (which is in $\mathcal{L}_{\text{uniq}}$, and hence also in $\mathcal{L}^+$) continues not to be valid in $\mathcal{T}$, while being valid in counterfactual structures. While C5 is not valid in $\mathcal{T}$, the generalization of C5 mentioned at the end of Section 2.1 is valid in $\mathcal{T}$ (Halpern, 2000), and is not valid in counterfactual structures. So the full class $\mathcal{T}$ of causal models is incomparable in expressive power to counterfactual structures with respect the language $\mathcal{L}^+$ (which is the language arguably most appropriate for $\mathcal{T}$). Formally, we have:

COROLLARY 3.8. *$\mathcal{T}$ is incomparable in expressive power to $\mathcal{M}$ with respect to the language $\mathcal{L}^+$.*

§4. Discussion. I have shown that the expressive power of causal models as models for counterfactuals is incomparable to that of the Lewis–Stalnaker "closest-world" possible-worlds semantics for counterfactuals; thus, the definition of counterfactuals in causal models is *not*, in general, compatible with the axioms of possible world semantics, although it is if we restrict to recursive causal models.

Specifically, causal models where the equations are recursive can be viewed as defining a strict subclass of the standard possible-worlds semantics. More precisely, a set of structural equation defines a world w (characterized by the unique solution to the equations) and can be implicitly viewed as defining an ordering relation on worlds such that, for every formula $\vec{X} = \vec{x}$, the solution to the equations when $\vec{X}$ is set to $\vec{x}$ determines a world $w_{\vec{X}=\vec{x}}$ that is the world closest to w according to the ordering such that $\vec{X} = \vec{x}$. Somewhat surprisingly, this is not the case if we go to the larger class of causal models defined by equations that are not recursive, but have a unique solution for all settings $\vec{X} = \vec{x}$. Of course, it is still the case that there is a world $w_{\vec{X}=\vec{x}}$ determined by the equations when $\vec{X}$ is set to $\vec{x}$. However, there is, in general, no ordering on worlds such that $w_{\vec{X}=\vec{x}}$ is the closest world to w according to the ordering. A closeness ordering on worlds places some restrictions (e.g., those characterized by the formula $\neg\varphi$ in Theorem 3.7) that do not hold in all causal models in $\mathcal{T}_{\text{uniq}}$.

So where does this leave us? It is still the case that, in causal models, a formula such as $\varphi \succeq \psi$ is true at a world w if ψ is true at some appropriate world w' satisfying φ. However, w' cannot be viewed as the "closest" world to w satisfying φ. This leaves open the question of whether there are other ways of defining "appropriateness" other than "closeness." I do not have strong intuitions here, but it is a question that is perhaps worth pursuing. My own feeling is that these arguments show that models in $\mathcal{T}_{\text{uniq}} - \mathcal{T}_{\text{rec}}$ are actually not good models for causality. It is quite difficult to verify that a nonrecursive causal model is in $\mathcal{T}_{\text{uniq}}$, as the model T given in Theorem 3.7 satisfying φ shows. I am not aware of any

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

interesting real-world situation that is captured by a model in $\mathcal{T}_{\text{uniq}} - \mathcal{T}_{\text{rec}}$. Further evidence of the "unreasonableness" of models in $\mathcal{T}_{\text{uniq}} - \mathcal{T}_{\text{rec}}$ is given by recent work of Zhang *et al.* (2012). They show that although the reversibility axiom blocks cycles of counterfactual dependence of length two, it does not block longer cycles. Indeed, they observe that the causal model T of Theorem 3.7 has a cycle of length three.

Does this mean that we should restrict to recursive models? There are certainly equations in physics (e.g., those connecting pressure and volume) that exhibit circular dependencies. Perhaps nonrecursive models would be appropriate for them (although once we add time to the picture, we may well be able to use a recursive model to capture any particular scenario). In a general nonrecursive model, there may be several solutions to an intervention (see Halpern & Pearl, 2005, Appendix A.4, for further discussion of this point). But this just corresponds to there being several worlds satisfying a formula φ that are closest to a given world, which is certainly allowed in Lewis's framework. I suspect that there is an interesting class of nonrecursive causal models that can be captured in Lewis's framework, and that these will turn out to be the models that actually arise in practice.

Zhang (2012) makes some progress on this issue. He proposes two condition on causal models, which he calls *solution-ful* and *solution-conservativeness*. The former condition is easy to understand: a causal model T is solution-ful if, for every context $\vec{u}$, the equations have a solution (not necessarily unique). The second condition is somewhat more complicated. T is solution-conservative if, for every context $\vec{u}$, if a solution to $T_{\vec{X}=\vec{x}}$ is consistent with $\vec{Y} = \vec{y}$, then every solution to $T_{\vec{X}=\vec{x} \wedge \vec{Y}=\vec{y}}$ is a solution to $T_{\vec{X}=\vec{x}}$. Zhang shows that a causal model satisfies these two conditions iff it satisfies analogues of all the axioms in one of Lewis's axiomatizations of causal counterfactuals (Lewis, 1973, p. 973). Thus, these two conditions are necessary for a causal model to be translatable to a counterfactual structure. They are not sufficient, since they are satisfied by all models in $\mathcal{T}_{\text{uniq}}$. There is clearly more to be done in understanding the connections between causal models and counterfactual structures.

Turning to a more technical point, it is worth trying to understand in more detail exactly what goes wrong in the Galles–Pearl argument. When making the comparison, Galles and Pearl did not use the axiom system AX. Rather, they used one of Lewis' axiomatizations of conditional logic from Lewis (1973). Nevertheless, the problem with their argument can be understood in terms of AX. Galles and Pearl show that all but one of the axioms in the system they consider follows from C3, C4, and C5. (They actually also implicitly use C1 and C2, but this is a minor point.) The remaining axiom is

$$((\varphi_1 \vee \varphi_2) \succeq \varphi_1) \vee ((\varphi_1 \vee \varphi_2) \succeq \varphi_2) \vee$$
$$[((\varphi_1 \vee \varphi_2) \succeq \psi) \Leftrightarrow ((\varphi_1 \succeq \psi) \vee (\varphi_2 \succeq \psi))].$$

For this axiom, they say "Because actions in causal models are restricted to conjunctions of literals that is, in the language of this paper, because in a formula in $\mathcal{L}_{\text{uniq}}$ of the form $\varphi \succeq \psi$, φ is a conjunction of formulas of the form $X = x]$, this axiom is irrelevant." They thus ignore the axiom. Unfortunately, this argument is flawed. If it were true, then it would be the case that the class of causal models in $\mathcal{T}_{\text{uniq}}$ would be *less* general than counterfactual structures since causal models would satisfy more axioms—all the relevant axioms satisfied by counterfactual structures, and, in addition, C5 (reversibility). However, as we have seen, the formula φ of Theorem 3.7 is satisfiable in $\mathcal{T}_{\text{uniq}}$, while $\neg\varphi$ is valid in M_f. The formula $\neg\varphi$ is equivalent to

$$([X_1 = 1](X_2 = 1 \wedge X_3 = 0) \wedge [X_2 = 1](X_3 = 1 \wedge X_1 = 0)$$
$$\Rightarrow \neg[X_3 = 1](X_1 = 1 \wedge X_2 = 0).$$

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

Thus, $\neg\varphi$ is a formula whose antecedent and conclusion are both formulas in $\mathcal{L}_{\text{uniq}}$. The argument for the validity of $\neg\varphi$ given in the proof of Theorem 3.7 is purely semantic, but, as I show in the appendix, $\neg\varphi$ can also be derived from AX together with V2 and V3. The derivation uses A4 in a crucial way. Note that A4 has disjunctions on the left-hand side of $\succeq$ and thus cannot be expressed in $\mathcal{L}_{\text{uniq}}$. But it is not irrelevant! Indeed, an easy argument (given in the appendix) shows that $\neg\varphi$ cannot be derived using AX $-$ {A4} together with V2 and V3 if we restrict to formulas in $\mathcal{L}_{\text{uniq}}$. Thus, we can start with assumptions in the language $\mathcal{L}_{\text{uniq}}$, end up with a conclusion in $\mathcal{L}_{\text{uniq}}$, but have a derivation that, along the way, uses A4 and has steps that involve formulas with disjunctions on the left-hand side of $\succeq$. The Galles and Pearl argument ignores this possibility.

§5. Acknowledgments. Thanks to Franz Huber, Judea Pearl, Jiji Zhang, and the anonymous reviewers of the paper for useful comments. A preliminary version of this paper appears in the Proceedings of the Twelfth International Conference on Principles of Knowledge Representation and Reasoning (KR 2010), 2010. Supported in part by NSF grants IIS-0534064, IIS-0812045, and IIS-0911036, and by AFOSR grants FA9550-08-1-0438 and FA9550-09-1-0266, and ARO grant W911NF-09-1-0281.

§6. Appendix: A derivation of $\neg\varphi$. In this appendix, I show that the formula

$$([X_1 = 1](X_2 = 1 \wedge X_3 = 0) \wedge [X_2 = 1](X_3 = 1 \wedge X_1 = 0) \Rightarrow \neg[X_3 = 1](X_1 = 1 \wedge X_2 = 0)$$

can be derived from AX together with V2 and V3. To simplify notation, I write $\vdash \varphi'$ if the formula φ' can be derived from AX, and $\vdash^{+} \varphi'$ if φ' can be derived from AX + {V2, V3}.

I first need a technical lemma.

LEMMA A.1. $\vdash (\varphi_1 \succeq \varphi_2 \wedge \varphi_2 \succeq \varphi_3) \Rightarrow (\varphi_1 \vee \varphi_2) \succeq \varphi_3$.

Proof. By A1, $\vdash \varphi_2 \succeq \varphi_2$, and by A4, $\varphi_1 \succeq \varphi_2 \wedge \varphi_2 \succeq \varphi_2 \Rightarrow (\varphi_1 \vee \varphi_2) \succeq \psi$. Thus,

$$\vdash (\varphi_1 \succeq \varphi_2) \Rightarrow (\varphi_1 \vee \varphi_2) \succeq \varphi_2. \tag{1}$$

Next observe that it easily follows from A1 and A4 that (a) $\vdash (\varphi_1 \wedge \neg\varphi_2) \succeq (\varphi_2 \Rightarrow \varphi_3)$ (since $(\varphi_1 \wedge \neg\varphi_2) \Rightarrow (\varphi_2 \Rightarrow \varphi_3)$ is a tautology) and (b) $\vdash (\varphi_2 \succeq \varphi_3) \Rightarrow \varphi_2 \succeq (\varphi_2 \Rightarrow \varphi_3)$ (since $\varphi_3 \Rightarrow (\varphi_2 \Rightarrow \varphi_3)$ is a tautology). By A4, RA1, and the observation that $((\varphi_1 \wedge \neg\varphi_2) \vee \varphi_2) \Leftrightarrow (\varphi_1 \vee \varphi_2)$ is a tautology, we can conclude that

$$\vdash (\varphi_2 \succeq \varphi_3) \Rightarrow ((\varphi_1 \vee \varphi_2) \succeq (\varphi_2 \Rightarrow \varphi_3)). \tag{2}$$

Let ψ be the formula $(\varphi_1 \succeq \varphi_2 \wedge \varphi_2 \succeq \varphi_3)$. From (1) and (2), it follows that

$$\vdash \psi \Rightarrow ((\varphi_1 \vee \varphi_2) \succeq \varphi_2) \wedge (\varphi_1 \vee \varphi_2) \succeq (\varphi_2 \Rightarrow \varphi_3)).$$

Applying A2 and RA1, it follows that

$$\vdash \psi \Rightarrow ((\varphi_1 \vee \varphi_2) \succeq \varphi_3,$$

as desired. □

Now the proof is easy. By Lemma A.1, it follows that

$$\vdash (X_1 = 1 \succeq X_2 = 1) \wedge (X_2 = 1 \succeq X_3 = 1) \Rightarrow (X_1 = 1 \vee X_2 = 1) \succeq X_3 = 1.$$

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

Applying the lemma again, we get that

$$\vdash (X_1 = 1 \vee X_2 = 1) \succeq X_3 = 1 \wedge (X_3 = 1 \succeq X_1 = 0) \Rightarrow (X_1 = 1 \vee X_2 = 1 \vee X_3 = 1) \succeq X_1 = 0.$$

Let ψ' be an abbreviation for $X_1 = 1 \vee X_2 = 1 \vee X_3 = 1$. Thus, $\vdash \varphi \Rightarrow (\psi' \succeq X_1 = 0)$. An analogous argument shows that $\vdash \varphi \Rightarrow \psi' \succeq X_2 = 0)$ and $\vdash \varphi \Rightarrow (\psi' \succeq X_3 = 0)$. By A2, we have that

$$\vdash \varphi \Rightarrow (\psi' \succeq (X_1 = 0 \wedge X_2 = 0 \wedge X_3 = 0). \tag{3}$$

By V2, we have that $\vdash^+ \psi' \Rightarrow (X_1 \neq 0 \vee X_2 \neq 0 \vee X_3 \neq 0)$. Thus, by A1 and RA2, we have that

$$\vdash^+ \varphi \Rightarrow (\psi' \succeq (X_1 \neq 0 \wedge X_2 \neq 0 \wedge X_3 \neq 0). \tag{4}$$

By A2, RA1, (3), and (4), we have that

$$\vdash^+ \varphi \Rightarrow (\psi' \succeq \mathit{false}).$$

By V3, it follows that $\vdash^+ \neg\varphi$, as desired. □

Note that all the axioms in AX other than A4 can be expressed using the language $\mathcal{L}_{\text{uniq}}$ (provided that the formulas φ and ψ mentioned in these axioms are taken to be conjunctions of primitive propositions or just single primitive propositions, as appropriate). It is easy to check that all these axioms are valid in $\mathcal{T}_{\text{uniq}}$, as are V2 and V3. Since φ is satisfiable in $\mathcal{T}_{\text{uniq}}$, $\neg\varphi$ cannot be proved from $(\text{AX} - \{\text{A4}\}) \cup \{\text{V2,V3}\}$.

BIBLIOGRAPHY

Bell, J. (1989). The logic of nonmonotonicity. *Artificial Intelligence*, **41**, 365–374.

Burgess, J. (1981). Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic*, **22**, 76–84.

Chellas, B. F. (1980). Basic conditional logic. *Journal of Philosophical Logic*, **4**, 133–153.

Friedman, N., & Halpern, J. Y. (1994). On the complexity of conditional logics. In *Principles of Knowledge Representation and Reasoning: Proc. Fourth International Conference (KR '94)*, Bonn, Germany. pp. 202–213.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundation of Science*, **3**(1), 151–182.

Grahne, G. (1991). Updates and counterfactuals. In *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*, Cambridge, MA. pp. 269–276.

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

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science*, **56**(4), 843–887.

Katsuno, H., & Satoh, K. (1991). A unified view of consequence relation, belief revision and conditional logic. In *Proc. Twelfth International Joint Conference on Artificial Intelligence (IJCAI '91)*, Sydney, Australia. pp. 406–412.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, **44**, 167–207.

Lewis, D. (1971). Completeness and decidability of three logics of counterfactual conditionals. *Theoria*, **37**(1), 74–85.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press

Lewis, D. K. (1974). Intensional logics without iterative axioms. *Journal of Philosophical Logic*, **3**, 457–466.
Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
Stalnaker, R. C. (1968). A theory of conditionals. In Rescher, N., editor. *Studies in Logical Theory*. Oxford, UK: Blackwell. pp. 98–112.
Zhang, J. (2012). A Lewisian logic of causal counterfactuals. *Minds and Machines*. To appear. doi: 10.1007/s11023-011-9261-z.
Zhang, J., Lam, W.-Y., & de Clercq, R. (2012). A peculiarity in Pearl's logic of interventionist counterfactuals. *Journal of Philosophical Logic*. To appear.

COMPUTER SCIENCE DEPARTMENT
CORNELL UNIVERSITY
ITHACA, NY 14853
E-mail: halpern@cs.cornell.edu

https://doi.org/10.1017/S1755020312000305 Published online by Cambridge University Press