Synthese
DOI 10.1007/s11229-016-1183-0

Causal counterfactuals are not interventionist counterfactuals

Tyrus Fisher[1]

Received: 5 February 2016 / Accepted: 26 July 2016

Abstract In this paper I present a limitation to what may be called strictly-interventionistic causal-model semantic theories for subjunctive conditionals. And I offer a line of response to Briggs' (Philos Stud 160:139–166, 2012) counterexample to Modus Ponens—given within a strictly-interventionistic framework—for the subjunctive conditional. The paper also contains some discussion of backtracking counterfactuals and backtracking interpretations. The limitation inherent to strict interventionism is brought out via a class of counterexamples. A causal-model semantics is strictly interventionistic just in case the procedure it gives for evaluating a subjunctive conditional requires making the values of the variables implicated in the antecedent independent from the values of the parents of these antecedent variables. Most causal-model semantic theories that have gained attention are strictly interventionistic.

Keywords Backtrackers · Counterfactuals · Causal models · Modus Ponens

1 Introduction

This paper presents a limitation inherent to what may be called *strictly interventionistic* semantic theories for subjunctive conditionals. The limitation is brought out via a class of counterexamples. These counterexamples are subjunctive conditionals that have intuitive readings on which they are false but for which there is no strictly-interventionistic model that falsifies them. The paper also contains some consideration of backtracking conditionals and backtracking interpretations, and it contains a line of response to Rachael Briggs' recently offered counterexample—within a strictly inter-

✉ Tyrus Fisher
tkfisher@ucdavis.edu

[1] 1240 Social Science and Humanities, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA

Published online: 08 August 2016

ventionistic framework—to Modus Ponens for the subjunctive conditional (Briggs 2012). A causal-model semantics is strictly interventionistic just in case the procedure it gives for evaluating a subjunctive conditional requires making the values of the variables implicated in the antecedent independent from the values of the parents of these antecedent variables. Most causal-model semantic theories that have gained attention are strictly interventionistic.[1]

A quick word about backtracking counterfactuals. To a first approximation a backtracker is a counterfactual such that some event implicated in its antecedent would occur later than some event implicated in its consequent.[2] Because strictly interventionistic semantic theories involve making antecedent variables independent of their parent variables, such theories do not handle backtracking counterfactuals well. Some philosophers and linguists view backtrackers as somehow non-standard and, so, outside the purview of a semantics for standard counterfactuals.[3] One motivation for this paper is to show that some features of a semantics that preclude adequate treatment of backtrackers also preclude adequate treatment of some non-backtrackers. Accordingly, whether the counterexamples I offer are themselves backtrackers will need to be considered. At least some of them are not backtrackers, as I will argue, but let me flag that if they were, the major premise in Briggs' counterexample to Modus Ponens would be a backtracker as well. I return to discussion of these issues in Subsect. 3.2. The difficulty for strict interventionism brought out in this paper should be appreciable regardless of one's views about the felicity or truth of backtrackers outside of special contexts.

The counterexamples alluded to above are strongly suggestive of the following adequacy condition for causal-model semantic theories: Appropriate evaluation of some *causal counterfactuals* requires that the values of the variables implicated in the antecedent of such conditionals remain sensitive (in a sense to be made more precise) to the values of their parent variables throughout the process of evaluating a counterfactual. By "causal counterfactual" I mean an *ontic counterfactual* that is about a causal system. We can, in the spirit of Lindström and Rabinowicz (1992), contrast ontic conditionals with epistemic conditionals. Epistemic conditionals are those properly evaluated by considering whether one should believe the consequent if one comes to believe (or were to come to believe) the antecedent. Ontic conditionals are those properly evaluated by considering how the world would be/have been if the antecedent obtains/obtained. I have no remarks about the nature of causation or causal systems except to say that I have in mind systems that are adequately represented by causal models like those characterized below.

Judea Pearl has given perhaps the most influential presentation of a strictly-interventionist semantics (Galles and Pearl 1998; Pearl 2000). Pearl gives a semantics for a fairly restricted language—one restricted to subjunctive conditionals containing

[1] Among strictly-interventionistic theories, Pearl's (2000) is probably the most influential. Eric Hiddleston's (2005) and Fisher's (2016) semantic theories are examples of causal-model theories that are not strictly interventionistic.

[2] In Subsect. 3.2, I consider several characterizations of backtrackers and whether my problem-case conditionals are backtrackers, or come out false only on a backtracking interpretation.

[3] See Lewis (1979) for the locus classicus of this view. See Downing (1959) for an early argument for the conclusion that all backtrackers are false.

only atomics or conjunctions of atomics in their antecedents and consequents. Joseph Halpern (2000) provides a richer strictly-interventionist semantics. It is richer in that it interprets subjunctive conditionals containing atomics or conjunctions of atomics in their antecedents and arbitrary Boolean constructions in their consequents. Rachael Briggs (2012) gives a strictly-interventionist semantics that is richer still. Briggs' semantics interprets subjunctive conditionals having arbitrary Boolean antecedents with consequents comprising arbitrary combinations of Boolean connectives and the subjunctive-conditional connective '$\Box\!\!\rightarrow$'—alternatively in this paper, '$>$'. So, Briggs' semantics is capable of interpreting subjunctive conditionals of the form: $\varphi > (\chi > \psi)$. It is at this level of syntactic complexity that the class of counterexamples I present emerges.

In Sect. 2, I present some guiding ideas behind strict interventionism and attempt to explain the attractiveness of this approach to interpreting subjunctive conditionals. In Sect. 3, I introduce Briggs' counterexample to Modus Ponens, and I argue that there is a wide class of right-embedded causal counterfactuals capable of coming out false but such that there is no strictly-interventionistic model that falsifies them. There too I offer some discussion of backtrackers and whether the counterexamples I give should be counted as such. In Sect. 4, I return to discussion of Briggs' counterexample to Modus Ponens for the subjunctive conditional and argue that in cases of the kind her counterexample derives from, a strictly-interventionist semantics delivers incorrect truth values for some subjunctive conditionals. And I use these problem-case conditionals to explain a way to resist the purported counterexample to Modus Ponens.

I've said that any strictly-interventionistic semantics will incorrectly evaluate a substantial class of conditionals. Here is a first example:

Match 1 I hold up a match and strike it, but it does not light. Then I say, "*If the match had lit, then (even) if it had not been struck, it would have lit*."

I submit that the conditional above is intuitively false. I do not mean that the conditional is false relative to all scenarios, worlds (or what have you) that it may be interpreted relative to.[4] Rather, I mean that it is intuitively false relative to some scenarios/worlds of the sort described in the minimal setup of *Match 1*. But, relative to any strictly-interventionist semantics the sentence above must come out true. Two points before moving on: First, for discussion of why some might be tempted to evaluate conditionals like *Match 1* as true, see Sect. 4 for discussion of the Import/Export principle. Second, should it be that the *Match 1* conditional is a valid sentence, then it constitutes a counterexample to standard Stalnaker/Lewis style semantic theories of counterfactuals.[5]

2 Interventionism and causal-model semantics

Some have thought that understanding causality can help us understand counterfactual dependence. Others, David Lewis prominently among them (Cf. Lewis 1973a, 2000),

[4] I use the word 'scenario' in addition to or instead of 'world' when I am concerned to talk in especially theory-neutral terms.

[5] E.g., the semantic theories laid out in Stalnaker (1968) and Lewis (1973b, 1979).

have endorsed some version of the converse thesis. Many who are attracted to a causal-model semantics for counterfactuals incline toward a view of the former sort. Such inclinations can make attractive the following sort of view about the truth conditions of subjunctive conditionals:

Causal truth, conditions(CTC) A subjunctive conditional $\varphi > \psi$ is true iff either ψ is independent of φ and true, or else φ is sufficient for bringing about ψ when holding fixed all the facts that are independent of φ,[6]

where these notions of *independence* and *bringing about* are causal in character for all or a substantial class of cases.

A guiding idea behind causal-model semantic theories for counterfactuals is that if we know well enough the details of a causal system—e.g., the states of its parts and their causal relations—such knowledge allows us to evaluate counterfactuals about that system. Causal models can be used to represent such details.

Leaving talk of causal systems behind in favor of talk about models of such things, we can take a causal model to be a triple $\mathcal{M} = \langle \mathcal{G} = \langle \mathcal{V}, \mathcal{R} \rangle, \mathcal{S}, \mathcal{A} \rangle$. $\mathcal{G}$ is a directed acyclic graph. $\mathcal{V}$ is a set of property or event-type variables, each of which takes values in a finite range $Ran(V)$. Often, $Ran(V) = \{0, 1\}$, where 0 and 1 represent no/yes, respectively. $\mathcal{R}$ is a set of edges, each of which represents a dependency relation and its directionality. And $\mathcal{S}$ is a set of sentences, each member of which further characterizes some members of $\mathcal{R}$ by describing the value of some member V in $\mathcal{V}$ or else how the admissible values of some V are constrained given the values of some Vs $\in \mathcal{V}$. Typically, the members of $\mathcal{S}$ are *structural equations* with some of the form $V = v$ (where $v \in Ran(V)$) and the rest of the form $V_i = f_i(V_j \ldots V_k)$, with '$f_i$' (denoting a function f_i) returning a value for V_i depending on the values of $V_j, \ldots, V_k$. Per standard terminology, let us say that a variable V' is a *parent* of a variable V just in case $\langle V', V \rangle$ is in $\mathcal{R}$. Conversely, V is a *child* of V' just in case $\langle V', V \rangle$ is in $\mathcal{R}$. A typical convention is to always place child/depender variables on the left side of equations and parent/dependees on the right. In this way the equations can be made to encode the structure and directionality of the edges. (I have not assumed this convention is in place. For this reason I have included the edges of the graph explicitly in the causal models.) Last, $\mathcal{A}$ is an assignment of values to the members of $\mathcal{V}$ that is consistent with $\mathcal{S}$.

Figuratively, to intervene on a causal system is to "reach in" from outside the system and set some part of it to a chosen state. To *intervene* on a causal model representing such a system is to set the values of some variables to some chosen values. An intervention on a variable V setting it to a value v isolates the pair from the variable-value-pairs that might otherwise influence the value that V takes. Because the pairs in $\mathcal{R}$ are indicated with arrows in pictorial representations of causal models, we can also say informally that V' is a parent of V just in case there is an arrow going out of V' and into V. The figures below contain examples of these target notions.

A crucial feature of strictly-interventionistic semantic theories is that when a variable V is intervened on, the value of V becomes independent of the values of its

[6] I've taken this "intuitive paraphrase" from Starr (2012); Starr cites Cumming (2009). See also Schulz's remarks on p. 241 of her "If you'd wiggled A, then B would have changed" (Schulz 2011).

parents. Informally, we can say that all the arrows going into V are "erased" and we can call such erasure, *pruning*.

Given truth conditions like (CTC), it is easy to see the prima facie plausibility of a strictly-interventionist semantics: (CTC) tells us that a subjunctive conditional $V = v > V' = v'$ is true if $V' = v'$ is an effect of $V = v$. Intervening on V in such a way as to set it to v allows us to see whether $V' = v'$ is among the effects of $V = v$. This suggests an interventionist recipe for evaluating a counterfactual $\varphi > \psi$:

(**IR**) (1) locate an accurate model $\mathcal{M}$ of the causal system that $\varphi > \psi$ is about; next,
(2) intervene on the variables mentioned in the antecedent φ;[7] then
(3) $\varphi > \psi$ is true iff the model obtained upon completion of (2) (per the rules of the particular semantics) satisfies ψ.[8]

I'll call this the interventionist recipe, (IR), and I'll take (IR) to constitute a core part of any interventionist causal-model semantics. A *strictly-interventionist semantics* is a causal-model semantics including (IR) and satisfying the condition:

(**SI**) When a variable V is intervened on so that it is made to take a value v, V remains set to v unless it is intervened upon again per an iterated application of (IR) (as in the case of an embedded conditional).[9]

How the interventionist uses causal models to interpret counterfactuals can now be illustrated with a well-worn example:

Firing Squad A prisoner D faces a firing squad. Executioners X and Y have rifles trained on the prisoner. As with X, Y fires iff the captain gives the order to fire. The captain gives the order to fire iff the judge delivers a death sentence. And each of X and Y are perfectly accurate and deadly. As it happens, the sentence is delivered and the prisoner is executed (Compare Pearl 2000, p. 207).

A causal model representing *Firing Squad* can be depicted as in Fig. 1.

Consider the counterfactual: *If the captain hadn't given the order, the prisoner would not have died* ($C = 0 > D = 0$). Our strict-interventionist recipe tells us to (1) locate an accurate causal model $\mathcal{M}$. Next, (2) intervene on C, setting its value to 0. This step isolates C from any variables that could otherwise influence the value it takes. This amounts to removing the sentence $C = J$ from the set of structural equations and replacing it with the sentence $C = 0$ and, with respect to the illustration of the graph, *pruning*—i.e., removing—all the arrows going into C. We then update the

[7] The presence of disjunctive antecedents is a complicating factor because there is more than one way to intervene in order to satisfy a disjunction. I have ignored this issue because it is orthogonal to my main points in this paper.

[8] If the semantics admits more than one model as relevant after the intervention, the analogue of (3) may, for example, be something like: $\varphi > \psi$ is true iff all the resulting appropriate submodels satisfy ψ (where what counts as an appropriate submodel is spelled out as part of the particular semantics).

[9] Other kinds of interventions are sometimes discussed. These are sometimes called "soft interventions". These soft interventions are such that intervened on variables are never completely isolated from the variables they would otherwise be causally related to. (Cf. Korb et al. 2004; Eberhardt and Scheines 2006). Such alternative notions of intervention have largely been ignored in the causal-model semantics literature to this point.

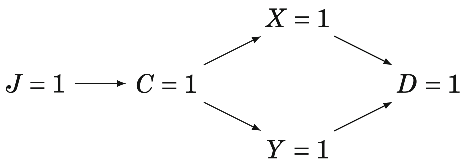

Fig. 1 *Firing Squad* $\mathcal{M} = \langle \mathcal{G} = \langle \mathcal{V}, \mathcal{R} \rangle, \mathcal{S}, \mathcal{A} \rangle$; with $\mathcal{V} = \{J,\ C,\ X,\ Y,\ D\ \}$, $\mathcal{S} = \{J = 1,\ C = J,\ X = C,\ Y = C,\ D = X \vee Y\}$, $\mathcal{A} = \{\langle J, 1\rangle,\ \langle C, 1\rangle,\ \langle Y, 1\rangle,\ \langle X, 1\rangle,\ \langle D, 1\rangle\}$. All variables are binary, taking 0 or 1. $\mathcal{R}$ is easily recoverable from the picture

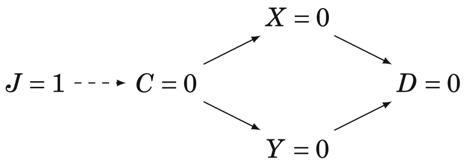

Fig. 2 The submodel $\mathcal{M}_{\neg C}$ of *Firing Squad* that results after intervening on C. $\mathcal{M}_{\neg C} = \langle \mathcal{G}' = \langle \mathcal{V}, \mathcal{R}' \rangle, \mathcal{S}', \mathcal{A}' \rangle$; with $\mathcal{V} = \{J,\ C,\ Y,\ X,\ D\ \}$, $\mathcal{S}' = \{J = 1,\ C = 0,\ X = C,\ Y = C,\ D = X \vee Y\}$, $\mathcal{A}' = \{\langle J, 1\rangle,\ \langle C, 0\rangle,\ \langle Y, 0\rangle,\ \langle X, 0\rangle,\ \langle D, 0\rangle\}$. All variables are binary, taking 0 or 1. $\mathcal{R}'$ is easily recoverable from the picture

value assignment, $\mathcal{A}$, to $\mathcal{A}'$, so as to preserve consistency with the new set of structual equations $\mathcal{S}'$. The resulting submodel $\mathcal{M}_{\neg C}$ is depicted in Fig. 2. Finally, we are in a position to perform (3). Since $\mathcal{M}_{\neg C}$ satisfies the consequent, a strictly-interventionist semantics tells us that the counterfactual $C = 0 > D = 0$ is true. This seems correct. In an effort to make what follows apply as widely as possible, I leave this sketch of a strictly-interventionist semantics at this level of generality. I have, for instance, left it open whether the members of $\mathcal{S}$ must be equations or whether they may express weaker relations, and whether $\mathcal{S}$ (or $\mathcal{S}'$) must determine a unique assignment $\mathcal{A}$ (or $\mathcal{A}'$).

3 Difficulty for strict interventionism

Using the *Firing Squad* case, Briggs (2012) has shown that a strictly-interventionist semantics that interprets right-embedded subjunctive conditionals invalidates Modus Ponens for the subjunctive conditional. On such a semantics,

$\mathcal{M} \models$ (a) $X > (\neg C > D)$
$\mathcal{M} \models$ (b) X
$\mathcal{M} \not\models$ (c) $\neg C > D$

In plain English the argument reads:

If X had fired, then (even) if the captain hadn't given the order, D would have died.
X fired.
If the captain hadn't given the order, D would have died.

To evaluate (a) on a strictly-interventionist semantics, we first intervene on X so as to satisfy (a)'s antecedent.[10] This yields the submodel $\mathcal{M}_X$ as depicted in Fig. 3.

[10] For a binary variable V taking 0 or 1 (for no/yes, respectively) I sometimes write 'V' as shorthand for '$V = 1$' and write '$\neg V$' as shorthand for '$V = 0$'.

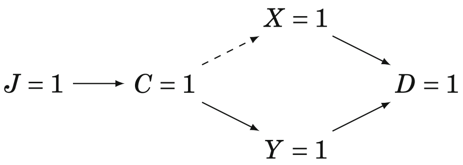

Fig. 3 *Firing Squad* $\mathcal{M}_X = \langle \mathcal{G}' = \langle \mathcal{V},\ \mathcal{R}' \rangle,\ \mathcal{S}',\ \mathcal{A} \rangle$; with $\mathcal{V} = \{J,\ C,\ Y,\ X,\ D\ \}$, $\mathcal{S}' = \{J = 1,\ C = J,\ X = 1,\ Y = C,\ D = X \vee Y\}$, $\mathcal{A} = \{\langle J, 1\rangle,\ \langle C, 1\rangle,\ \langle X, 1\rangle,\ \langle Y, 1\rangle,\ \langle D, 1\rangle\}$. All variables are binary, taking 0 or 1. $\mathcal{R}'$ is easily recoverable from the picture

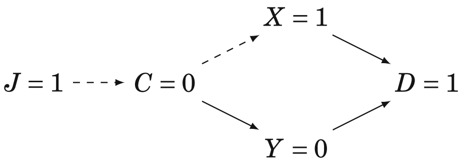

Fig. 4 *Firing Squad* $M_{X_{\neg C}} = \langle \mathcal{G}'' = \langle \mathcal{V},\ \mathcal{R}'' \rangle,\ \mathcal{S}'',\ \mathcal{A}' \rangle$; with $\mathcal{V} = \{J,\ C,\ Y,\ X,\ D\ \}$, $\mathcal{S}'' = \{J = 1,\ C = 0,\ X = 1,\ Y = C,\ D = X \vee Y\}$, $\mathcal{A}' = \{\langle J, 1\rangle,\ \langle C, 0\rangle,\ \langle X, 1\rangle,\ \langle Y, 0\rangle,\ \langle D, 1\rangle\}$. All variables are binary, taking 0 or 1. $\mathcal{R}''$ is easily recoverable from the picture

Next we check whether $\mathcal{M}_X$ satisfies (a)'s consequent, $\neg C > D$. Since this consequent is itself a subjunctive conditional, we intervene again, setting C to 0. This yields the submodel $\mathcal{M}_{X_{\neg C}}$ (Fig. 4). Indeed, $\mathcal{M}_{X_{\neg C}}$ satisfies D. So $\mathcal{M} \vDash$ (a) $X > (\neg C > D)$. It is easy to verify that $\mathcal{M}$ satisfies X but does not satisfy $\neg C > D$. Such is Briggs' counterexample (Cf. Briggs 2012, p. 150). And, I confess, I think the counterexample has some intuitive attractiveness. I return to discussion of Brigg's counterexample in Sect. 3.

With a strictly-interventionistic evaluation procedure before us and a clear way to apply it to embedded conditionals, I turn to considering a class of counterexamples to strict interventionism.

Match 2 I hold up a dry well-made match. I strike it, and it lights. Then I say "*If the match were struck and it lit, then if it hadn't been struck, it would have lit*". $(Strike = 1 \wedge Light = 1) > (Strike = 0 > Light = 1)$

I think $(Strike = 1 \wedge Light = 1) > (Strike = 0 > Light = 1)$ is intuitively false relative to Match 2. And I think that if the model of Fig. 5 is an appropriate model for the case, then if the match weren't struck, it wouldn't have lit. Otherwise, ours must be a world with some other available and salient cause of match lighting that the model fails to represent; for if it lights, there must be some cause of the lighting. All this holds whether such scenarios are counterfactual or otherwise. That is, even in scenarios where the match is alight, if the cause of the lighting had been absent, it wouldn't have lit.

I have used the words "appropriate" and "salient". So let me explain. It may well be that which causal relations are relevant to the evaluation of a counterfactual is sensitive to features of context and, accordingly, that which model is appropriate for evaluation of a counterfactual is sensitive to context. So by "salient cause" I mean causes that are relevant in this sense. By "appropriate" model I mean a model that is accurate with considerations of salience taken into account. But in Match 2 if there

$$Strike = 1 \longrightarrow Light = 1$$

Fig. 5 *Match 2* $\mathcal{M}_{M2} = \langle \mathcal{G} = \langle \mathcal{V}, \mathcal{R} \rangle,\ \mathcal{S},\ \mathcal{A} \rangle$; with $\mathcal{V} = \{Strike,\ Light\}$, $\mathcal{S} = \{Strike = 1,\ Light = Strike\}$, ($\mathcal{R}$ and $\mathcal{A}$ are easily recoverable from the picture). Both variables are binary, taking 0 or 1

$$Strike = 0 \dashrightarrow Light = 1$$

Fig. 6 *Match 2.* $\mathcal{M}_{M2_{\neg Strike}} = \langle \mathcal{G}' = \langle \mathcal{V}, \mathcal{R}' \rangle,\ \mathcal{S}',\ \mathcal{A}' \rangle$; with $\mathcal{V} = \{Strike,\ Light\}$, $\mathcal{S}' = \{Strike = 0,\ Light = 1\}$, ($\mathcal{R}' = \emptyset$, and $\mathcal{A}'$ is easily recoverable.) Both variables are binary, taking 0 or 1

is some salient possible cause of the match lighting apart from its being struck, then this possible cause should be depicted in the model. Analogously, in the the *Firing Squad* case if there is some salient possible cause of X's pulling the trigger other than the Captain's order, then this possible cause should be depicted in the model, else the model is simply inaccurate in an important way and hence inappropriate for evaluating the conditional at issue.

Any semantics that directs us to hold fixed the value of the antecedent variables, thereby making them independent of the values of their parents, will misevaluate cases like Match 2 relative to contexts such that $\mathcal{M}_{M2}$ of Fig. 5 is an appropriate model for the case. To illustrate: a strictly-interventionist semantics will have us look first to $\mathcal{M}_{M2}$ and then, after intervening on *Strike* and *Light*, have us intervene again on *Strike* and look to $\mathcal{M}_{M2_{\neg Strike}}$, as depicted in Fig. 6. Hence, $(Strike = 1 \wedge Light = 1) > (Strike = 0 > Light = 1)$ comes out true, as I claim it should not.

To appropriately interpret $(Strike = 1 \wedge Light = 1) > (Strike = 0 > Light = 1)$, we need a semantics that allows *Light* to remain causally sensitive to its parent throughout the evaluation procedure. At this point the adequacy condition at issue can be made more precise.

Adequacy condition A causal model semantics for counterfactuals should admit cases in which the variables implicated in the antecedent of a counterfactual remain *causally sensitive* to their parents throughout the evaluation procedure;

where, to say that the antecedent variables remain causally sensitive to their parents is to say that the dependency relations between the antecedent-variables and their parents remain unchanged throughout the evaluation procedure. And it is easy to see that to accept this adequacy condition is nearly just to accept the rejection of strict interventionism.

Could a modified version of interventionism handle problem cases like that described in *Match 2*? Yes; consider the principle (SCI):

Side-constrained interventionism(SCI) When evaluating a conditional relative to a causal model $\mathcal{M}$, perform a non-trivial intervention just in case $\mathcal{M}$ does not satisfy the antecedent of the conditional. (Where a trivial intervention is one that results in no changes to $\mathcal{M}$.)

The idea is that an (SCI) semantics would be a semantics such that the (SCI) principle informs our understanding of (2) in (IR). So an (SCI) semantics is still a strictly-interventionistic semantics, constrained though it is.

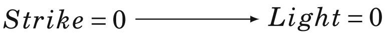

Fig. 7 *Match 2* $\mathcal{M}_{M2_{\neg Strike'}} = \langle \mathcal{G} = \langle \mathcal{V}, \mathcal{R} \rangle,\ \mathcal{S},\ \mathcal{A} \rangle$; with $\mathcal{V} = \{Strike,\ Light\}$, $\mathcal{S} = \{Strike = 0,\ Light = Strike\}$, ($\mathcal{R}$ and $\mathcal{A}$ are easily recoverable). Both variables are binary, taking 0 or 1

I think there is reason for the causal-model semanticist to be attracted to an interventionist semantics constrained per (SCI). Non-trivially intervening on variables already satisfying the antecedent at issue can seem like an overly-legalistic application of the guiding idea behind interventionism.

In *Match 2*, because the values of *Strike* and *Light* equal 1 from the start, on an (SCI) semantics there is no call to non-trivially intervene on *Light*. On such a semantics, one would look first to the model depicted in Fig. 5 and (non-trivially) intervene only on *Strike*. This yields the model depicted in Fig. 7, which falsifies the consequent and makes the conditional come out false, as it should.

An (SCI) semantics will validate Modus Ponens. Such a semantics satisfies a centering principle. In this context, centering can be taken to be the principle that if $\mathcal{M} \vDash \varphi$, then $\mathcal{M} \vDash \varphi > \psi$ iff $\mathcal{M} \vDash \psi$. To see that centering holds on an (SCI) semantics suppose we are given a counterfactual $\varphi > \psi$ to be evaluated relative to a given $\mathcal{M}$. Suppose further $\mathcal{M} \vDash \varphi$. Then the set of $\mathcal{M}$'s variables to be intervened on in the course of evaluating $\varphi > \psi$ is empty and the model resulting from the trivial intervention is just $\mathcal{M}$. Hence, $\mathcal{M} \vDash \varphi > \psi$, iff $\mathcal{M} \vDash \psi$.

So far I've presented some conditionals that are capable of coming out false under natural circumstances but for which no strictly-interventionist semantics admits a falsifying model. The trick has been to construct a right-embedded conditional $\varphi > (\chi > \varphi)$ such that χ describes an event that tells causally against an occurrence of an event described by φ. I've also pointed out that (SCI), a fairly conservative means of resolving the particular problem case discussed in this section has the effect that Modus Ponens is validated by the resulting semantics.

3.1 Counterfactual problem cases for strict interventionism

One might hope that adopting (SCI) would suffice to resolve the counterexamples associated with right-embedded conditionals and (SI). But this is not so; there are plenty of cases involving subjunctive conditionals with *false* antecedents such that intervening on the variables implicated by the leftmost antecedent guarantees the wrong evaluation. The recipe for constructing such problem cases is nearly the same as for those above. The case given at the outset of this paper is an appropriate example:

***Match 1* (again)** I hold up a match and strike it, but it does not light. Next I say, "*If the match had lit, then (even) if it had not been struck, it would have lit.*" $Light = 1 > (Strike = 0 > Light = 1)$

In *Match 1*, the background story given is just as with *Match 2* save that the match did not light, making the antecedent false. A causal model for *Match 1* is depicted in Fig. 8. Because the *Match 1* story makes it certain that striking the match is not sufficient for its lighting, an error variable E is included in the causal model for *Match 1*, where E represents the combined influences of all other factors causally relevant to

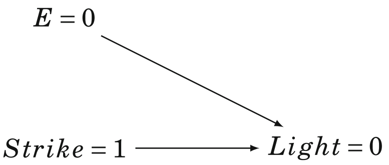

Fig. 8 *Match 1* $\mathcal{M}_{M1_{Strike}} = \langle \mathcal{G} = \langle \mathcal{V}, \mathcal{R}\rangle, \ \mathcal{S}, \ \mathcal{A}\rangle$; with $\mathcal{V} = \{Strike, \ Light, \ E\}$, $\mathcal{S} = \{Strike = 1, \ Light = Strike \times E\}$, $(\mathcal{R} = \langle Strike, \ Light\rangle, \ \langle E, \ Light\rangle$, with $\mathcal{A}$ easily recoverable). Variables are binary, taking 0 or 1

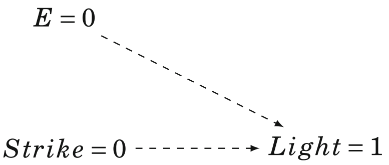

Fig. 9 *Match 1* $\mathcal{M}_{M1_{\neg Strike}} = \langle \mathcal{G}' = \langle \mathcal{V}, \mathcal{R}'\rangle, \ \mathcal{S}', \ \mathcal{A}'\rangle$; with $\mathcal{V} = \{Strike, \ Light, \ E\}$, $\mathcal{S}' = \{Strike = 0, \ Light = 1\}$, $(\mathcal{R}' = \emptyset$ with $\mathcal{A}'$ easily recoverable). Variables are binary, taking 0 or 1

whether the match lights. Evaluating *Match 1* per SCI yields the submodel of Fig. 9. The *Match 1* conditional comes out true relative to (SCI), and this seems incorrect.

Since *Match 1* is a case involving a false antecedent where, on a straightforward reading of the conditional and relative to a natural context, intervening per strict interventionism yields the wrong truth value, we get the result that adding (SCI) to an otherwise strictly-interventionist semantics is insufficient for resolving the problem. So an adequate causal-model semantics must depart more substantially from strict-interventionism than (SCI) guarantees.

How might one resist what I've said against strict interventionism? One suggestion is that something about the conditionals I've offered encourages them to be read non-ontically.[11] The obvious alternative is to try to read the embedding connective or the embedded connective as an epistemic conditional. If this suggestion is right, then the offered counterexamples fail, since a semantics should not be faulted for inappropriately interpreting sentences it's not meant to.[12]

I think the suggestion is ultimately uncompelling: Consider the conditional "If the headlamp were emitting light, then if I had blinked, the headlamp would not have been emitting light". We are confident that this conditional is false because of the causal irrelevance of my eye-lid movements to the event implicated by the right-embedded consequent. The conditional is most straightforwardly read as thoroughly ontic. But I can't see why such a conditional should suddenly suggest a non-ontic reading if we substitute the right-embedded consequent "I had blinked" for something such as "the headlamp were to lack batteries". Such a substitution yields another counterexample of the sort I have been offering:

[11] Thanks to Rachael Briggs for this point.

[12] There is controversy concerning the availability of an epistemic reading for subjunctive conditionals. See, e.g., Edgington (1995), Veltman (2005), Schulz (2007), Kratzer (2012), and Khoo (2015).

Headlamp I hold up a headlamp in good working condition. I say "*If the headlamp were emitting light, then if it had had no batteries, the headlamp would be emitting light*". $Emit = 1 > (Batt = 0 > Emit = 1)$

The suggestion that either connective here is amenable to an epistemic reading ultimately seems implausible to me especially because the presence or absence of batteries in a headlamp is causally relevant to the event implicated by the antecedent. Similar remarks apply to the match-striking cases presented above.

3.2 Is the problem just backtracking?

A different worry about what I've said against strict interventionism concerns whether the problem conditionals I have presented are false because they are backtrackers or are otherwise being evaluated relative to a backtracking interpretation. If so, then what I've said would be much less interesting, since it is well known that strictly interventionistic accounts do not handle backtrackers well. In this subsection I'll say a bit about how others have understood the notion of a backtracker and the related but distinct notion of a *backtracking interpretation*. And I will argue that the problem conditionals I have offered are not backtrackers and are false on their non-backtracking interpretations.

It will be handy from here on to have an intuitive characterization of what a backtracking interpretation is. Justin Khoo has a "rough intuitive gloss" that should serve well in this respect (Khoo forthcoming, p. 3). Following Khoo, we can, at the intuitive level, characterize the notion of a non-backtracking interpretation of a subjunctive conditional and its complementary notion of a backtracking interpretation as follows. On a "non-backtracking interpretation …one 'punches' its antecedent-event into the causal history of the world and then plays things out from there to see whether its consequent is thereby made true" (Khoo forthcoming, p. 3). On the other hand, when evaluating a counterfactual according to a backtracking interpretation "one does a bit of 'detective work' to figure out in what circumstances its antecedent would have been true, and then, making the requisite changes to history to bring about its antecedent, plays things out accordingly to see whether its consequent is thereby made true" (Khoo forthcoming, p. 3).

Before proceeding, let me highlight that if either the *Match 1* conditional or the *Headlamp* conditional are backtrackers or come out false only on a backtracking interpretation, then so is/does the conditional $X > (\neg C > D)$ (If X had fired, then (even) if the captain hadn't given the order, D would have died). This last conditional figures crucially in Briggs' counterexample to Modus Ponens. So if $X > (\neg C > D)$ is false only on a backtracking interpretation, we already have good reason to reject the counterexample to Modus Ponens and the semantics that yields it. This is because the counterexample to Modus Ponens would then be the result of using a semantics to interpret sentences of a sort we know it is ill-suited to interpret. But, for reasons given below, I don't think any of these conditionals are backtrackers or come out false on their non-backtracking interpretions.

As Jonathan Bennett points out, discussion of backtracking seems to start with P. B. Downing's paper "Subjunctive Conditionals, Time Order, and Causation" (Downing

1959; Bennett 1974, p. 391).[13] Downing tells us that a "backward looking subjunctive conditional" is a conditional such that "the consequent contains references to times earlier than any times referred to in the antecedent" (Downing 1959, p. 130). In reviewing Lewis' *Counterfactuals*, Bennett discusses some of Lewis' independently arrived at ideas about backtracking, citing Downing's paper in the process and thereby calling Lewis' attention to the Downing paper (Bennett 1974, p. 391; Bennett (2003), p. 205). Bennett introduced the term 'back-tracking' (apparently in his 1974) (Bennett 2003, p. 208).[14] But it is in Lewis' subsequent discussion in "Counterfactual Dependence and Time's Arrow" that we first find the phrase 'back-tracking counterfactual'. The "Time's Arrow" paper has since influenced most subsequent discussion of backtrackers.

In addition to Downing's characterization of backtracking counterfactuals, two more characterizations worth considering are that of Lewis (1979) and that of the linguist Anna Arregui (2005). First Lewis' characterization:

> A counterfactual saying that the past would be different if the present were somehow different may come out true under the special resolution of its vagueness, but false under the standard resolution. If so call it a *back-tracking counterfactual* (Lewis 1979).

I want to highlight two features of Lewis' characterization.

First, on Lewis' characterization only counterfactuals about the past are backtrackers—his characterization is formulated with reference to *past* and *present* times.[15] But it is not just counterfactuals about the past that seem relevant to discussions about backtracking. For example, the conditional, "If Heidi were to win the lottery next year, she'd have first acquired a ticket". There is no reference to the past or present in this conditional since all the possible events mentioned would occur in the future. Given Lewis' purposes and the intended scope of his discussion in "Times Arrow", the restricted nature of his characterization may not be a defect, but it is worth noting that his characterization is substantially restricted in its applicability.

Secondly, built into Lewis' characterization is that to be a backtracker a counterfactual must come out false under the "standard resolution". This falsity condition decides too much of what is at issue in many discussions of backtrackers, as often much of what is at issue is whether some backtrackers come out true on a standard resolution.

Arregui's characterization is more inclusive. She writes:

> *Real backtrackers* are conditionals that explicitly claim that if a certain hypothesis held at time *t* (past present or future), something different would have

[13] Downing distinguishes between backward looking subjunctive conditionals, forward looking conditionals, and simultaneous conditionals (Downing 1959). Some of my problem-case conditionals count as simultaneous conditionals. That is, conditionals "with consequents containing references to times as early as the earliest time referred to in the antecedent" (Downing 1959, p. 136).

[14] Bennett writes, "My story involves *backtracking*—counterfactualizing back in time and forward again. (That is the sense I gave the term when introducing it into the literature. Others have used it for just the first half of that" (Bennett 2003, p. 208).

[15] Thanks to an anonymous referee for pointing out that Lewis' discussion is restricted in this way.

> happened at some earlier time t' (Arregui 2005, p. 84; italics and parentheses in original).

Arregui's characterization does not share those features of Lewis' highlighted above. However, a feature of Arrequi's characterization—and many others—to worry about is that for a conditional to count as a backtracker, its antecedent and consequent must describe ways things might be in such a way that particular times are picked out.[16] But very often our sentences do not describe things in such a way that any particular times are picked out by them.[17] This is a limitation of many characterizations of backtrackers and one I only wish to note.

I have chosen three characterizations that I think give a sense of how backtrackers are typically characterized by philosophers and linguists (and suggested that it is hard to unproblematically say just what a backtracker is). With Downing's, Lewis', and Arregui's characterizations in mind now, consider the following conditionals, with parenthetical phrases added to encourage a particular, perfectly accessible, reading:

(δ') I hold up a match and strike it, but it does not light. Then I say, "*if the match had lit, then if it had been extremely windy (at the time of the lighting), then it would (still) have lit*".

Headlamp$'$ *If the headlamp were emitting light, then if it had no batteries (at the time it would have been emitting light), then the headlamp would be emitting light.*

Both of these conditionals pose the same trouble for strict interventionism as the *Match 1* conditional. But in both of these conditionals no event occurring prior to the event implicated in the antecedent is described or need otherwise be implicated. Further, there is no suggestion that if the present were different the past must have been different. Rather, both counterfactuals tell us at most that some point or period in time would be different in some respect if that very point or period of time had been different in another respect. So neither (δ') or *Headlamp*$'$ is a backward looking conditional, a back-tracking conditional, or a real backtracker (in the senses of Downing, Lewis, or Arregui, respectively).

The more serious worry, however, is that even if the problem conditionals I have offered are not backtracking conditionals, on their false readings they are nonetheless being evaluated relative to a backtracking interpretation. That is, the worry is that to come out false they require an evaluator to go in for backtracking reasoning and so are not being resolved in the standard non-backtracking way.

That some conditionals which are not explicitly backtrackers are nonetheless evaluated relative to a backtracking interpretation is nicely illustrated by Downing's original case (1959) [later popularized by Lewis (1979, p. 456) and Bennett (2003, p. 205)]:

> Jim and Jack quarrelled yesterday, and Jack is still hopping mad. We conclude that (1) if Jim asked Jack for for help today, Jack would not help him. But wait: Jim is a prideful fellow. He never would ask Jack for help after such a quarrel;

[16] Downing's characterization induces effectively the same worry as it requires that the antecedent contain "references" to times earlier than that of the antecedent.

[17] Thanks to Adam Sennet for this point.

> (2) if Jim were to ask Jack for help today, there would have to have been no quarrel yesterday. In that case Jack would be his usual generous self. So (3) if Jim asked Jack for help today, Jack would help him after all (Lewis 1979, p. 456); numbers in parentheses added).

While (2) is explicitly a backtracking counterfactual and (3) is not, (3) is intuitively false *unless* one goes in for the backtracking reasoning/detective work that supports (2). This is evidenced by our initial inclination to accept (1) prior to being taken through the backtracking reasoning.

For the "it's just backtracking" objection to my counterexamples to succeed, it must be that when considering the problem-case conditionals, in order to get the false reading we entertain changes to history that are not properly entertained on a non-backtracking interpretation. I'll try to show now that my problem cases should come out false even on a non-backtracking interpretation.

Consider *Match 1* again:

***Match 1* (again)** I hold up a match and strike it, but it does not light. I say "If the match had lit, then if it had not been struck it would have lit".

In order to get clear on the relevant possible interpretations of *Match 1*, let us syntactically mark them. Let a subscripted 'B' indicate a backtracking interpretation of a subjunctive conditional and let a lack of subscript indicate a non-backtracking interpretation.[18]

a. $L >_B (\neg S > L)$
b. $L > (\neg S >_B L)$
c. $L > (\neg S > L)$

So the "it's just backtracking" response can now clearly be stated as the worry that when conditionals such as *Match 1* come out intuitively false, we are reading them as of types a. or b., but if we are careful to give them their type c. reading, they are true.[19]

Can it be shown that when one intuitively evaluates *Match 1* as false this is not because she is reading it per a. or b.? I do not know. That is, if someone reports he is inclined to read *Match 1* in the manner given by a. or b. and that this is why he judges *Match 1* to be false, I do not know how to show him he is not really doing that. But I don't need to show that. Notice that for the "it's just backtracking" response to succeed it would not be enough that readings of *Match 1* per a. or b. are intuitive and yield the value False. Rather, for the response to succeed, *Match 1* must also come out true on its c. reading. And, as I'll try to show, there is good theory-neutral reason for judging that *Match 1* comes out false even on its c. reading. So let us focus on a reading of *Match 1* that is stipulated to be non-backtrackng (i.e., of type c.) and what can be discerned of such a reading.

Bennett and Lewis have shown us there is good reason to think that when we intuitively evaluate subjunctive conditionals about the past on a non-backtracking

[18] Thanks to an anonymous reviewer for suggesting the use of this notational device.

[19] Of course one could also try to go in for a reading of *Match 1* as of types a. *and* b., i.e.: $L >_B (\neg S >_B L)$. But I think the point I need can be made without explicitly taxonomizing the above conditional as of a distinct kind.

interpretation we do not—and should not—hold the past entirely fixed until exactly the time of the antecedent. Bennett gives a case that seems to show that if the ordinary way of evaluating subjunctive conditionals had us holding things fixed right up until the time of the antecedent, then lots of intuitively false conditionals would be intuitively true. Here is Bennett's case:

> The dam burst at 8:47 p.m., and within two minutes the waters had swept through the valley, killing nine occupants of cars on the valley road. Reflecting on the gratifying fact that the dam-burst did little other serious harm, when it might have been expected to kill thousands, someone remarks 'If there had been no cars on the road just then, no lives would have been lost' (Bennett 2003, p. 210).

The asserted conditional looks true. But if the correct way to evaluate subjunctive conditionals on a non-backtracking interpretation is to hold things fixed until exactly the time of the antecedent, then, Bennet points out, the following should also be true:

> If there had been no cars on the road just then, we would be investigating the mystery of where they had all gone to.
>
> If there had been no cars on the road just then, this would be evidence of a miraculous divine intervention—the sudden removal of the cars at the very moment when the dam burst (Bennett 2003, p. 210).

Both of the above are intuitively false. So, Bennett concludes that "A normal competent speaker who asserts No cars > No-deaths envisages a state of affairs in which it *smoothly* comes about that the valley lacks cars when the dam bursts" (Bennett 2003, p. 210).

What does Bennett mean by 'smoothly'? I think Bennett cannot subject his language to too much analysis on pain of making his point no longer applicable to intuitive evaluations of subjunctive conditionals. Too much analysis would see us arrive at the level of explicit semantic theorizing. Still, Bennett is able to elaborate a fair bit:

> (1) The facts about our use of counterfactuals show that when we think $A > C$ we do not in general envisage A as becoming true abruptly and discontinuously, with a notable bump …(2) The facts also show that we do not envisage A's becoming true out of a past that may be greatly different from [the actual past]. Putting those two together: the facts about our use of counterfactuals show that we ordinarily think of A as fairly smoothly becoming true out of a past pretty much like the actual world's, which means that (3) they show that we take our counterfactuals to require forks from [actual-world]-like worlds. This is about as far as we can go in describing the frame of mind of some-one—not a theorist of conditionals—who thinks $A > C$ [Bennett (2003, p. 225); brackets added].[20]

Lewis agrees with Bennett that a good non-backtracking analysis of counterfactuals needs to include a "transition period" of counterfactual divergence from how things actually were in order to "provide an orderly transition from actual past to counterfactual present or future" (Lewis 1979, p. 463). Lewis' reasons here are similar to Bennett's: Concerning a match that was not struck, Lewis remarks

[20] 'Fork' and 'bump' are technicals terms for Bennett. See Bennett (2003, pp. 202–221).

> Right up to t, the match was stationary and a foot away from the striking surface. If it had been struck at t, would it have travelled a foot in no time at all? No; we should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future" (Lewis 1979, p. 463).[21]

Taking these remarks of Bennett's and Lewis' to heart we can use Khoo's characterization (still at a rough intuitive level) to arrive at an evaluation of *Match 1* on the interpretation corresponding to c. We begin by "smoothly punching in" that the match lit. To do this, we consider ways things might have been that are exactly as things actually were until shortly before the striking event, but we must allow for a transition period during which things would have been slightly different just prior to the antecedent time(s) so that the match lights. (Presumably, we should be thinking of ways such that whatever conditions prevented the actual strike from being sufficient for lighting no longer succeed in stymying the match's ignition.) Next, we evaluate the embedded conditional relative to our set of antecedent-satisfying scenarios. So we "smoothly punch in" that the match was not struck. Does the match light? If we take Bennett's and Lewis' arguments to heart, it does not.[22]

The match doesn't light because with respect to the *Match 1* problem-case, there are no other salient possible causes of match-lighting available on any way of smoothly departing from the actual past in the moments just before the lighting event. Any such emergent cause would constitute a rather abrupt and discontinuous departure (or too big of a miracle in Lewis' manner of speaking). And all this holds on an evaluation at the intuitive level given by an interpretation that was stipulated to be non-backtracking.

Of course, the meanings of 'punch in' and 'smooth' can stand to be made much more precise. But any attempt at such further precision will see us leaving the level of intuitive evaluation behind. At which point, we will simply be investigating how our explicitly-theoretical semantic accounts of non-backtrackers would have us evaluate *Match 1*. We already know what the theory-driven judgments look like here: A strict-interventionist theory will see all the conditionals I have offered come out true. While Lewis' (1979), Bennett's (2003), and a number of other theories will see them come out false.[23]

It is because strictly interventionistic theories guarantee that variables implicated in the antecedent of a conditional become independent of their parents that such theories do not handle backtrackers appropriately. It is for this same reason that the counterexamples I have produced pose trouble for these theories. Importantly, the counterexamples I have offered come out false relative to normal-enough contexts

[21] Lewis makes these remarks in the course of motivating an analysis he will reject—his Analysis 1—but he stands by the lesson of the quoted remarks.

[22] But notice also that even if one rejects Bennett's smoothness requirement and Lewis' transition period, it looks like if one insists on evaluating *Headlamp*, for example, by unsmoothly punching in that the headlamp has no batteries at precisely the time of the embedded antecedent it seems to me there is good reason to conclude that the lack of batteries would stymie the headlamp's capacity for emitting light. Similar remarks hold with respect to the match's ignition and the extreme wind referenced in (δ).

[23] Other semantic theories that see the conditionals at issue come out false include Hiddleston's (2005), causal-model theory, Fisher's (2016) causal-model theory, and the theory of Khoo (2016).

and on non-backtracking interpretations. But there is no way for them to come out false on a strictly-interventionistic theory.

4 A counterexample to Modus Ponens?

I have yet to diagnose why Briggs' counterexample to Modus Ponens might seem attractive and explain why I think the pull is to be resisted. The answers, I suggest, have to do with the prima-facie attractiveness of the so-called *Import/Export* rule. Let '$\rightarrow$' name a conditional connective of interest. Then Import/Export says that a sentence of the form $(\varphi \wedge \chi) \rightarrow \psi$ is equivalent to the corresponding sentence of the form $\varphi \rightarrow (\chi \rightarrow \psi)$. By *Exportation* in what follows I will mean the rule corresponding to the left-to-right direction of entailment. Import/Export is valid for the material conditional, of course, but can be seen to fail for the subjunctive conditional. Still, the apparent equivalences that the principle purports to warrant are often quite attractive. I suggest that the intuitive appeal of Brigg's counterexample to Modus Ponens is a result of reading (a) $X > (\neg C > D)$ as equivalent to (or at least as entailed by) $(X \wedge \neg C) > D$. But we should not do this.

Some will have long since noticed the structural similarity between Briggs' counterexample to Modus Ponens and the counterexamples to Modus Ponens for the indicative conditional offered by McGee (1985).[24] Both Briggs' counterexample and McGee's involve right-embedded conditionals. Here I want to discuss what McGee (1985) has to say about Modus Ponens and the subjunctive conditional. This is because of how the rule of Exportation figures in McGee's consideration of the Subjunctive conditional and in the diagnosis I am working toward as to why Briggs' counterexample to Modus Ponens has the attractiveness it does.

McGee argues that Modus Ponens for the subjunctive conditional should be rejected because Exportation is intuitively valid and because, as he shows, if one endorses both Modus Ponens and Exportation (and a rather weak conditional-proof principle) for the subjunctive conditional, one is committed to the logical indistinguishability of the subjunctive and material conditional (1985, p. 465). But no acceptable semantics for the subjunctive conditional will yield such indistinguishability.

McGee's indistinguishability argument can be given as follows: Propositional logic gives us

(1) $(A \supset B) \wedge A \vDash B$

Now, assuming entailment is a sufficently strong relation to underwrite the truth of a subjunctive conditional (i.e., if $\varphi \vDash \psi$ then $\vDash \varphi > \psi$), we get (2).

(2) $[(A \supset B) \wedge A] > B$.

But then by Exportation for the subjunctive conditional we get as a logical truth:

(3) $(A \supset B) > (A > B)$.

By Modus Ponens for the subjunctive conditional, assuming $A \supset B$ allows us to derive $A > B$. Hence,

[24] Briggs notes the similarity in her footnote 8 (Briggs 2012, p. 150).

(4) $(A \supset B) \supset (A > B)$

For the other direction of the equivalence, assume $A > B$. Further, assume A. Then by Modus Ponens we get B. So we get $A \supset B$ under the scope of our initial assumption. And so we have:

(5) $(A > B) \supset (A \supset B)$

Therefore, endorsing both Modus Ponens and Exportation for the subjunctive conditional commits one to the the subjunctive and material conditional being equivalent.[25]

McGee notes that he does not have a counterexample to Modus Ponens for the subjunctive conditional, but he judges that there is good inductive evidence for the validity of Exportation for the subjunctive conditional (1985, p. 466). He thus concludes that it is Modus Ponens we should give up. There are, however, counterexamples to unrestricted Exportation for the subjunctive conditional. Here is one:

N1 If Nelson were a philosopher and not a philosopher, then he would be a philosopher ($[P \wedge \neg P] > P$). True.

N2 If Nelson were a philosopher, then if he were not a philosopher, he would be a philosopher ($P > [\neg P > P]$). False.

The consequent of N1 is (relevantly) entailed by its antecedent. But, turning attention to N2, for any way things could be such that Nelson is a philosopher, it is false that if he were not a philosopher he would (still) be a philosopher. So N2 is false.

Might there be a restricted version of Exportation for the subjunctive that *is* valid? Briggs' (2012) semantics validates a restricted Import/Export principle: Let bolded Roman letters $\mathbf{A}, \mathbf{B}, \ldots$ range over conjunctions of atomic formulas "in which no two conjuncts mention the same variable" (Briggs 2012, p. 150). Then, "where **A** and **B** share none of their variables, $(\mathbf{A} \wedge \mathbf{B}) \mathrel{\Box\!\!\rightarrow} C$ is equivalent to $\mathbf{A} \mathrel{\Box\!\!\rightarrow} (\mathbf{B} \mathrel{\Box\!\!\rightarrow} C)$" (Briggs 2012, p. 156).

But, there are intuitive counterexamples to this restricted Import/Export principle. To illustrate, suppose I strike a match on a calm, windless day, consequently it lights. Now consider:

(δ) If the match had lit, then if it had been extremely windy the match would have lit. $Light = 1 > (Wind = 1 > Light = 1)$

(ϵ) If the match had lit and it was extremely windy, then the match would have lit. $(Light = 1 \wedge Wind = 1) > Light = 1$

(δ) is false—if it had been extremely windy, the match might not have lit. But (ϵ) is true—the antecedent entails the consequent. So the restricted Import/Export principle is intuitively invalid.

It might be thought that one way to resist this conclusion is to hold that ontic subjunctive conditionals have two available readings, and the difference between these readings emerges only with embedded conditionals. On one of these readings, the thought continues, Restricted Import/Export is valid. As discussed below, I do think that we often treat utterances of right-embedded conditionals as equivalent to their

[25] Much the same argument was given by Gibbard (1981) and later by Edgington (2014, p. 2.5).

Import/Export counterparts, and rightly so. However, in order to impact the logic(s) of subjunctive conditionals, the different readings at issue would have to constitute a semantic ambiguity.[26] By contrast, the hypothesis I explore some paragraphs below involves different readings emerging at the level of pragmatics.

In any case, we can use the notational device and the remarks of Khoo, Bennet, and Lewis examined in 2.2. to help put aside the worry about ambiguity. The version of Restricted Import/Export at issue is that which would apply to (δ) on its non-backtracking interpretation. But owing to the considerations in 2.2, we can see that (δ) is false on this interpretation. But since the antecedent of (ϵ) relevantly entails its consequent, (ϵ) is true.

Notice now that (δ) and (ϵ) are syntactic analogues of some conditionals involved with Briggs' counterexample to Modus Ponens. (δ) is an analogue of (d) and (ϵ) is an analogue of (e).

(d) $X = 1 > (C = 0 > X = 1)$.
(e) $(X = 1 \wedge C = 0) > X = 1$.

(δ) is false because the consequent has us consider a counterfactual scenario in which an event tells causally against the event described by the right-embedded consequent. By similar reasoning, (d) should be judged false. But (e) is true.

If we are agreed that (d) $X = 1 > (C = 0 > X = 1)$ is false, I believe we should be agreed that (a) $X > (\neg C > D)$ is false. This is because, given the model at issue in the *Firing Squad* case, (a) looks good only if one accepts that the counterfactual scenario implicated by $\neg C$, when embedded in (a), is a scenario including a gunshot from X; but if one denies (d), one should deny just that.

To better see what I think has gone wrong, consider the question: "If X had fired, then if the captain hadn't given the order, would X have still fired?" That is, should $X > (\neg C > X)$ be judged true, or false? I think the reasoning appropriate for answering this question looks like this: Among the things known about the case is that X did in fact fire and that, according to the model, he fired *because* the captain gave the order. So, our question becomes: "If X had fired, as he did, would he have fired if the actual cause of his firing were absent?" Since the model also shows that X's firing was not overdetermined, the answer to this question should be "no". I wish to emphasize this last point. If the model is an appropriate and sufficiently accurate representation of the relevant causal system, then our world is such that the captain's order is the only salient cause of X's firing. Hence, *If X had fired, then (even) if the captain hadn't given the order, X would have fired*, i.e., $X > (\neg C > X)$, is false. But a strictly-interventionist semantics tells us it is true.

But what explains why it is often so easy to miss the difference between a subjunctive conditional and its Import or Export counterpart? In trying to answer this question, I can do little more than tell a story about why this is so. But I think it is not a bad story. Perhaps our readiness to go in for an imported reading is due in part to the

[26] Thanks to an anonymous reviewer for highlighting the possibility of multiple semantic interpretations for the subjunctive conditional. Ambiguity theses for counterfactuals are not unheard of. See, for example, Kvart (1994). For a recent development of an ambiguity hypothesis to do with counterfactuals and casual models see Lee (2015).

cognitive load of counterfactual reasoning—when we import we need only make a single *imaginative leap* to a counterfactual scenario, whereas when we do not go in for the imported reading we must make multiple imaginative leaps, one for each antecedent we meet as the conditional is parsed.

By "imaginative leap" I mean whatever cognitive task we work to complete when we consider a set of nearest (i.e., most similar or otherwise most relevant) counterfactual ways for things to be. The greater difficulty associated with evaluating right-embedded counterfactuals may arise because for each embedded antecedent an evaluator must make a new imaginative leap—i.e., consider a new set of nearest counterfactual ways for things to be. And with each such leap the evaluator must keep track of whether and how the newly considered antecedent should bear on the occurrence of events implicated by an antecedent further to the left.

But importation works fine in many cases. This is because it quite often works out that if this were an *A* world/scenario, then the nearest *B* world/scenario would be an *A* and *B* world/scenario. The cases for which Import/Export leads us astray are just those such that an event described in an embedded antecedent would affect an event described by an antecedent embedded further left or right. Reasoning out how such antecedents might bear on each other is often hard, and there is often no payoff for this work over simply going in for the imported reading. This is meant to explain why we often go in for imported readings of conditionals and why such a practice is a reasonable heuristic.

In the presence of a background body of theory concerning counterfactual reasoning, what I have just said is suggestive of some predictions about our assessment of right embedded conditionals. Here is one: If what I have suggested is correct, quite apart from any difficulties to do with assessing the equivalence or non-equivalence of Import/Export counterparts, we should expect multiply-right-embedded conditionals to often be harder to understand than their Export counterparts.[27] Indeed, this is borne out. Consider:

> If Kripke were at the party [if Strawson stayed home, (if Anscombe were there, then Malcolm would have been there)].

To my mind that is a difficult sentence to understand (even more so without the parentheses). Now consider the imported counterpart:

> If Kripke were there and Strawson stayed home and Anscombe were there, then Malcolm would have been there.

The imported counterpart is much easier to understand. Plausibly, I think, this is because it tells me explicitly which conditions the counterfactual scenario at issue must satisfy. The non-Import counterpart on the other hand requires that I reason out the features of the kind of counterfactual scenario at issue, keeping track of how the

[27] Notice that this is not the point, highlighted by no-truth-value (NTV) theorists such as Gibbard (Gibbard 1981), that left embedded indicatives are often hard to interpret. As noted by Sennet and Weisberg, right-nested conditionals are typically not thought to exhibit the phenomenon to do with left nested conditionals that the arguments of the NTV theorists involve (Sennet and Weisberg 2011, p. 450).

antecedents bear on each other as I move through the reasoning (unless I just go in for the imported reading of course).[28]

But what of the cases where we don't go in for Importation, what explains that? I have in mind examples like: *If X had fired, then if the Captain hadn't given the order, X would have (still) fired* [$X > (\neg C > X)$]. I think many of us, perhaps after some thought, are inclined to judge this conditional false. Hence, we are not applying Importation. Grice's maxims may provide some insight here. Importing on $X > (\neg C > X)$ yields a logical truth, so to understand an assertion of such a sentence in this way would have us attributing to the speaker a violation of the maxims of relation and quantity.[29]

Consideration of Grice's maxims may also reveal something more about cases where we do apply Importation: Given the background provided by the *Firing Squad* case and with respect to the conditional $X > (\neg C > D)$, if we go in for the import, an assertion of "X fired" in the antecedent remains relevant to our judgment of whether the consequent "D died" is satisfied in the relevant counterfactual scenario(s). But if we understand the conditional in the manner that I suggest gives its proper semantic content, we find that an utterance of $X > (\neg C > D)$ is such that the outermost antecedent becomes irrelevant to evaluating the conditional. On such a reading, if the captain doesn't give the order, D won't die. And that's regardless of whether X fired in the scenario called up by the outermost antecedent. Thus, the non-imported reading—though semantically proper—runs afoul of the maxim of relation. However, if we do go in for the Import reading, X's firing remains relevant to our evaluation of the consequent. Throughout this paper I've been arguing that the non-imported reading is the correct one, and I have tried to adduce considerations above to bring this out. Nevertheless, I appreciate that in some cases the non-imported reading may engender a feeling of infelicity. So I have offered an explanation of that infelicity.

There are often good pragmatic reasons for going in for the imported reading. But this does not yield semantic equivalence for the sentence types at issue. And I believe that once sufficient attention is paid to the difference between the Import/Export counterparts at issue in Briggs' counterexample, the counterexample becomes unconvincing.

5 Concluding remarks

I have argued that for sufficiently syntactically-complex subjunctive conditionals, strictly-interventionistic semantic theories admit of a certain class of counterexamples. These counterexamples are conditionals of the form $\varphi > (\chi > \varphi)$, where χ describes an event that tells causally against the occurrence of the event described by φ. It is because Briggs' causal-model semantics is strictly-interventionistic that it generates a counterexample to Modus Ponens for the subjunctive conditional. Accord-

[28] Thanks to an anonymous reviewer for urging me to express my points here more clearly and for suggesting that I consider predictions that the story here might yield.

[29] Recall that the maxim of relation says: try to be relevant and say things that are pertinent. The maxim of quantity says: try to be as informative as possible and give as much information as needed but no more (Grice 1975).

ingly, I have argued that in some cases, including the case that generates the apparent counterexample to Modus Ponens, a strictly-interventionistic semantics will evaluate some subjunctives incorrectly and that once the correct evaluations are appreciated, the counterexample to Modus Ponens become unconvincing. Appropriate causal-model evaluation of many counterfactuals—even many non-backtrackers—appears to require that the variables implicated by the antecedent remain sensitive to the values of variables that are causally upstream. But strict-interventionism, even modified per (SCI), makes this impossible. At present, the literature on causal-model semantic theories has little to offer in the way of non-strictly-interventionistic causal-model accounts. One aim of this paper has been to encourage the production and exploration of such accounts.

Acknowledgments Thanks to Bas van Fraassen and Ted Shear for their helpful comments on an earlier draft. Thanks also to two helpful anonymous reviewers. Most of all, much thanks and gratitude are due to my adviser, Adam Sennet, for all his help.

References

Arregui, A. (2005). *On the accessibility of possible worlds: The role of tense and aspect*. Doctoral dissertation, University of Massachusetts, Amherst.

Bennett, J. (1974). Counterfactuals and possible worlds. *Canadian Journal of Philosophy*, *4*(2), 381–402.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Clarendon Press.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*, 139–166.

Cumming, S. (2009). On what counterfactuals depend. Unpublished.

Downing, P. B. (1958–1959). Subunctive conditionals, time order, and causation. *Proceedings of the Aristotelian Society*, *59*, 125–140.

Eberhardt, F., & Scheines, R. (2006). Interventions and causal inference. In *Research Showcase @ CMU*.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Edgington, D. (2014). Indicative conditionals. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Stanford: Stanford University.

Fisher, T. (2016) Counterlegal dependence and causation's arrows: Causal models for backtrackers and counterlegals. *Synthese*. doi:10.1007/s11229-016-1189-7.

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

Gibbard, A. (1981). Two recent theories of conditionals. In W. L. Harper, R. Stalnaker, & C. Pearce (Eds.), *Ifs*. Dordrecht: Reidel.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Speech acts* (pp. 41–58). New York: Academic Press.

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

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39*(4), 632–657.

Khoo, J. (2015). On indicative and subjunctive conditionals. *Philosopher's Imprint*, *15*(32), 1–40.

Khoo, J. (2016). Backtracking counterfactuals, revisited. *Mind forthcoming*.

Korb, K., Hope, L., Nicholson, A., & Axnick, K. (2004). Varieties of causal intervention. In C. Zhang, H. W. Guesgen, & W.-K. Yeap (Eds.), *PRICAI 2004: Trends in Artificial Intelligence* (Vol. 3157, pp. 322–331). Lecture Notes in Computer Science. Berlin: Springer.

Kratzer, A. (2012). *Modals and conditionals, Volume 36 of Oxford studies in theoretical linguistics*. Oxford: Oxford University Press.

Kvart, I. (1994). Counterfactuals: Ambiguities, true premises, and knowledge. *Synthese*, *100*(1), 133–164.

Lee, K. Y. (2015). Causal models and the ambiguity of counterfactuals. In W. van der Hoek, W. Holliday, & W. F. Wang (Eds.), *Logic, rationality, and interaction* (Vol. 5). Berlin: Springer.

Lewis, D. (1973a). Causation. *Journal of Philosophy*, *70*, 556–567.

Lewis, D. (1973b). *Counterfactuals*. Oxford: Blackwell Publishing.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*, 455–476.

Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, *97*(4), 182–197.

Lindström, S., & Rabinowicz, W. (1992). The ramsey test revisited. *Theoria*, *58*(2–3), 131–182.

McGee, V. (1985). A counterexample to Modus Ponens. *The Journal of Philosophy*, *82*(9), 462–471.

Pearl, J. (2000). *Causality: Models, reasoning, and inferences* (1st ed.). Cambridge: Cambridge University Press.

Schulz, K. (2007). *Minimal models in semantics and pragmatics: Free choice, exhaustitivity, and conditionals*. Ph.d thesis, University of Amsterdam, Amsterdam.

Schulz, K. (2011). If you'd wiggled A, then B would've changed. *Synthese*, *179*(2), 239–251.

Sennet, A., & Weisberg, J. (2011). Embedding if and only if. *Journal of Philosophical Logic*, *41*, 449–460.

Stalnaker, R. C. (1968). A theory of conditionals. In R. Stalnaker (Ed.), *Studies in logical theory* (pp. 98–112). Oxford: Blackwell Publishing.

Starr, W. (2012). Subjunctive conditionals and structural equations: Unpublished. Retrieved from williamstarr.net.

Veltman, F. (2005). Making counterfactual assumptions. *Journal of Semantics*, *22*, 159–180.