Synthese (2017) 194:4983–5003
DOI 10.1007/s11229-016-1189-7

# Counterlegal dependence and causation's arrows: causal models for backtrackers and counterlegals

**Tyrus Fisher**[1]

Received: 14 December 2015 / Accepted: 31 July 2016 / Published online: 19 August 2016


**Abstract** A *counterlegal* is a counterfactual conditional containing an antecedent that is inconsistent with some set of laws. A *backtracker* is a counterfactual that tells us how things would be at a time earlier than that of its antecedent, were the antecedent to obtain. Typically, theories that evaluate counterlegals appropriately don't evaluate backtrackers properly, and vice versa. Two cases in point: Lewis' (Noûs 13:455–476, 1979a) ordering semantics handles counterlegals well but not backtrackers. Hiddleston's (Noûs 39(4):632–657, 2005) causal-model semantics nicely handles backtrackers but not counterlegals. Taking Hiddleston's account as a starting point, I offer steps toward a theory capable of handling both counterlegals and backtrackers. The core contribution of this paper is a means for evaluating counterlegals relative to *minimally-illegal models*.

**Keywords** Counterfactuals · Causal models · Backtracking counterfactuals · Counterlegal conditionals · Conditionals

## 1 Introduction

A *counterlegal* is a counterfactual conditional containing an antecedent that is inconsistent with some set of laws. By 'law' I mean a sentence describing a relationship between event or property types that is invariant over the possible events or prop-

✉ Tyrus Fisher
tkfisher@ucdavis.edu

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

erty instantiations admitted by a linguistic context.[1] A *backtracker* is a counterfactual that tells us how things would be at a time earlier than that of its antecedent, were the antecedent to obtain. Some philosophers view counterlegals or backtrackers as deviants beyond the scope of ordinary treatments of counterfactuals. I think this is a mistake, not least because counterlegals and backtrackers often figure in good reasoning that seems abnormal in no important way.

Typically, theories that handle counterlegals well handle backtrackers poorly, and vice versa. Two cases in point: Lewis' (1979a) ordering semantics handles counterlegals well enough but not so for backtrackers. Hiddleston's (2005) causal-model semantics handles backtrackers well but evaluates all counterlegals as vacuously true. Taking Hiddleston's account as a starting point, I offer steps toward a single theory capable of appropriately handling both counterlegals and backtrackers. The counterlegals I am concerned with in this paper are those containing in their antecedents a description of some combination of events that is inconsistent with the laws of the given context. The core contribution of this paper is a means for evaluating such counterlegals in terms of *minimally-illegal models*. A precise characterization of minimal illegality is offered below, but the rough idea is that a model is minimally-illegal when it represents a minimal violation of a causal system's laws while satisfying the antecedent of a counterlegal and maximally preserving the model's original assignment of values to its variables.

In the "Time's Arrow" paper, Lewis (1979a) argues that there is a "standard resolution" of the vagueness that "infects" the antecedent of counterfactuals and that on this standard resolution (excluding cases to do with "time travel or the like") the future depends counterfactually on its past, but not vice versa (Lewis 1979a, p. 457). On Lewis' view, backtrackers (as I have characterized them) are true on their standard resolution only if they have a true consequent (Lewis 1979a, b, pp. 457–458). Such a view can make it seem reasonable to be untroubled by backtracking conditionals that are not adequately handled by one's favorite semantic theory. This is because, it may be claimed, the favored theory is a semantic theory only for counterfactuals as standardly resolved, and backtrackers (with false consequents) require a non-standard resolution for their truth.

Counterlegals are another sort of counterfacual that some philosophers have claimed are, in some sense, beyond the purview of a semantics for ordinary counterfactuals. Bennett, for example, holds that counterlegals are "beyond the reach" of our ability to interpret counterfactuals, and, hence, may be ignored when assessing the virtues of a semantics for counterfactuals intended to interpret counterfactuals of the sort that "crop up in everyday life" (2003, p. 228).[2] Moreover, Bennett tells us that, with the notable exception of Pollock's (1976, 1981) work, there is no widely-known and detailed work

[1] I take it that such sentences often include those that are law-like according to a favored scientific theory. Consequently, I take it that the project at hand bears on the problem of interpreting counterfactuals containing antecedents that are nomically impossible according to a favored scientific theory. I do not, however, have feelings of attachment to any traditional, weighty conception of law-likeness from the philosophy of science of past decades. I am, for example, happy to understand talk of scientificlaw-likeness in terms of invariance across possible manipulations, where "invariance across possible manipulations" is understood along the lines of a Woodward (2003)-style account.

[2] Some of Bennett's remarks about counterlegals are reproduced at the conclusion of this paper.

on counterlegals (Bennett 2003, p. 228). This paper is motivated by the view that a good theory of counterfactuals should handle backtrackers and counterlegals as well as other sorts of counterfactuals.

The paper proceeds as follows. In Sect. 1 I discuss backtrackers and their status as problem cases for Lewis' theory. In Sect. 2 I introduce Hiddleston's causal-model semantics (TC), discuss counterlegals, and take care to show that the failure of TC to allow non-trivial evaluation of counterlegals is a limitation of that theory. These first two sections are intended to demonstrate the desirability of a semantics for both backtrackers and counterlegals and suggest something of the difficulty associated with producing such a semantics. Let me highlight the general difficulty here: To non-trivially evaluate counterlegals, a semantic theory needs some way of non-trivially interpreting antecedents that are inconsistent with the laws of a context. But when backtrackers come out true, this is often because the laws of the context dictate that some event(s) prior to the times implicated by the antecedent would need to play out (or have played out) differently in order for the antecedent to be satisfied. So, roughly, evaluation of counterlegals demands a procedure for breaking laws, but evaluating backtrackers requires holding them fixed. In Sect. 3 I consider what a good causal-model semantics for backtrackers and counterlegals should look like. There I define *minimal illegality* and use this definition to construct a causal-model semantics, CTC, intended to interpret both backtrackers and counterlegals. In Sect. 4 I quickly highlight a contrast between CTC, Hiddleston's TC, and Pearl-style (2000) interventionist causal-model semantic theories. And I close by considering how the proposed semantics stands in relation to some of Bennett's (2003) remarks about counterlegals and the prospects for an adequate semantics for them.

## 2 Backtrackers and Lewis' semantics

As I use the term, a backtracker is simply a counterfactual with an antecedent describing an event that takes place at a later time than some event described in its consequent.[3] Typically, a backtracker purports to tell us how the past would be different, were its antecedent to obtain. If one understands 'backtracking counterfactual' in the way I suggest, then according to Lewis (1979a) if we confine our attention to the standard resolution of the vagueness of counterfactuals (and ignore cases involving "time travel or the like"), then a backtracker is true just in case its consequent is true (Cf. Lewis 1979a, b, p. 458).[4] But plenty of back-

[3] Characterizations of what it is to be a backtracker typically suffer from lack of precision in order to avoid falsity. The characterization I give is no exception. The clauses of conditionals do not typically describe events so precisely as to single out a unique time or period of time at which the implicated events must (have) occur(red). So, there is often no such thing as *the* time of the event described in the antecedent or the consequent. Thanks to Adam Sennet for this point. Nevertheless, there are backtracking counterfactuals and I hope my characterization points well enough to the paradigm cases.

[4] Lewis (1979a) appears to think that the falsity of backtrackers with false consequents is delivered by his similarity metric (reproduced in Sect. 1 below) for evaluating counterfactuals that he offers in the "Times Arrow" paper. And indeed the similarity-metric does typically weight similarity of earlier events much more than later ones. Still, Lewis' belief (subject to caveats to do with time travel cases and the like) that a backtracker is true on its standard resolution just in case its consequent is true seems to be largely motivated

trackers with false consequents are as clearly true as any counterfactual. Here are two:

(a) If that infant were eight months older, then she would have been conceived at some time earlier than she actually was.
(b) If Scotty were in Australia now, then he would not have joined us (in California) for dinner earlier this evening.

In "Counterfactual Dependence and Time's Arrow", Lewis (1979a) amends his (1973) theory in such a way that it explicitly treats backtracking counterfactuals as deviants. The "skeleton" of Lewis' theory remains the same:

> A counterfactual $\varphi > \psi$ is (non-vacuously) true iff some (accessible) world where both $\varphi$ and $\psi$ is true is more similiar to the actual world than any world where $\varphi$ is true and $\psi$ is false (Compare Lewis 1979a, p. 465).

As Lewis tells us, if his theory is to facilitate evaluation of particular counterfactuals, this skeleton must be fleshed out with details about the operative similarity metric. Indeed, this is a lesson we should take from the well-known apparent counterexamples to Lewis' theory (which I will not rehearse here) given to us by, among others, Fine (1975) and Tichý (1976). To this end, Lewis gives us a system of weights designed to specify the appropriate similarity relation. These weights have the effect that similarity between worlds during times prior to the antecedent counts for much more than similarity after. Here is Lewis' often-cited system of weights:

(1) It is of the first importance to avoid big, widespread diverse violations of law.
(2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
(3) It is of the third importance to avoid even small, localized, simple violations of law.
(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly (Lewis 1979a, p. 472).

If the above metric is used to determine the world(s) most like ours among those satisfying the antecedent of interest, true backtrackers very often come out false. Here is one that comes with a story:

> ***Tracking Photons*** A photon is fired through a vacuum from Point $A$ to Point $B$ at time $t_1$. The photon arrives at $B$ at $t_3$. Photons travel as fast as anything can, so *if the photon had arrived at* $B$ at $t_2$ *(where* $t_1 < t_2 < t_3$*) it would have been fired at some time earlier than* $t_1$.

If Lewis' system of weights is used to determine the world(s) most like ours among those satisfying the antecedent of interest, we get the wrong result. To see this, consider

Footnote 4 continued
by his belief that past events do not counterfactually depend on future events. It is not too hard to see that holding this view can make it attractive to think that any counterfactual saying that an earlier time would be the same were some later time different will be true, while any counterfactual saying that some earlier time would be different were some later time different will be false.

two worlds, $w_1$ and $w_2$, each satisfying the antecedent of interest. Let $w_1$ be a closest world that matches our world perfectly until just before the time picked out by the antecedent and also such that the photon (miraculously) arrives at $B$ at $t_2$. Let $w_2$ be a closest world satisfying the antecedent such that the photon is fired at some time earlier than $t_1$. The photon's behavior at $w_1$ is miraculous to be sure, but a single photon breaking a law at a moment in a small region of space is a microphysical event that looks to be a small, local miracle. So both worlds accord with condition (1). But $w_1$ better accords with Lewis' (2). And the worlds do the same, or near enough, with respect to (3). Hence, $w_1$ is closer than $w_2$. So there is a world falsifying the conditional at issue that is closer than any of those that might otherwise verify it. Wrong result.

Lewis' semantics for counterfactuals yields a nice logic. And, when enough is said about the operative similarity metric, it returns apparently correct truth values for an impressive array of particular counterfactuals. In this section I've sought only to point out that some backtrackers are true and that the more-or-less standard Lewis-style semantics leaves room for improvement with respect to its treatment of backtrackers. Next I consider a rather different account, which excels in some ways that Lewis' does not.

## 3 Hiddleston's causal theory of counterfactuals

The last few decades have seen a number of authors advance semantic analyses of counterfactuals that make use of causal models (e.g., Galles and Pearl 1998; Pearl 2000; Halpern 2000; Hiddleston 2005; Briggs 2012). Eric Hiddleston's (2005) Theory of Counterfactuals, TC, is one such semantics. A causal model has the structure of a directed acyclic graph. The edges of these graphs are intended to represent causal-dependency relations among the event types or properties modeled by its nodes. One can use such a model to represent a collection of event types or properties plus some of the causal relations that those things stand in to each other.

A guiding idea behind accounts such as Hiddleston's is that if we understand well enough the causal dependency relations that characterize a causal system, then we can utilize that understanding to determine the truth or falsity of counterfactuals about events of the types that the system concerns. The models Hiddleston utilizes are, he tells us, "stripped down versions of the 'causal models' of Pearl (1999, 2000) and Spirtes, Glymour and Scheines (SGS 2000)" (Hiddleston 2005, p. 638).[5] Such a model is a triple $\langle G, E, A\rangle$. $G$ is a finite directed acyclic graph having event or property variables $V_1, \ldots, V_n$ as its nodes.[6] $E$ is a set of sentences describing the relations between variable-value pairs. These sentences represent the laws (or invariances if you prefer)

[5] They are "stripped down" in that they are free of various constraints imposed by Pearl and SGS: The models need not satisfy either the Markov or Faithfulness conditions. And the sentences in $E$ may describe relations that are probabilistic and/or non-functional. Pearl (2000) characterizes the Causal Markov Condition on page 30 (see also page 19) and faithfulness (there called 'stability') on page 48.

[6] The restriction to finite graphs is to ensure that, given a model $M$, the set of $\varphi$-minimally altered models (as defined below) relative to $M$ will always be non-empty. Thanks to Alex Kocurek for highlighting that without appropriate constraints on the class of models, some causal models will fail to have a non-empty set of $\varphi$-minimally altered models.

of the causal system represented by the model. For the purposes of this paper, these members of $E$ are material conditionals in our metalanguage. But because they are held fixed across ranges of possible assignments to the variables of the models, they have an import much like strict conditionals. These sentences are to be thought of as describing relations among variable-value pairs, and these relations are allowed to be non-functional and/or probabilistic. $A$ is an assignment of values to the nodes that satisfies the relations described by the members of $E$.[7]

On Hiddleston's account, a counterfactual $\varphi > \psi$ is true in a model $M$ just when $\psi$ is true in all $\varphi$-models that are *minimally altered* with respect to $M$.[8]

With an eye toward explicating Hiddleston's (2005) account, let us look at a particular counterfactual considered by Hiddleston that Lewis' account has difficulty with and that TC handles nicely.

> ***Betting on Coin Tosses*** Alice offers Ben a bet on a fair, indeterministic coin toss. Ben bets tails. Alice flips the coin, and it lands heads. Ben loses. Alice says to Ben, "If you had bet heads, you would have won" [Bet-Heads > Win]. That seems true (Hiddleston 2005, p. 634).

On Lewis' semantics, since the probability that the coin comes up heads and the probability that the coin comes up tails are each 0.5, some of the nearest worlds verify the consequent and some don't. So Bet-Heads > Win comes out false, as it should not.

In contrast, on Hiddleston's account we proceed as follows. Given the details of the story, we first locate a *good model* of the causal system it describes. Hiddleston directs us to call a set of events a *case* and to think of an event as an object's (or ordered set of objects) having a property. TC tells us that a model is a *good model* for a case $C$ iff

a. the properties represented in $M$ are instantiated by objects in $C$,
b. $M$'s law set $E$ contains laws that are accurate of $C$, or accurate enough, and
c. $M$ is complete enough to accurately represent the causal relations among the events of $C$ that appear in $M$ (Compare Hiddleston 2005, p. 648).

To evaluate Bet-Heads > Win we'll look to $M = \langle G, E, A\rangle$, where, $G = \langle \boldsymbol{V}, R\rangle$ such that,

$$
\begin{aligned}
&\boldsymbol{V} = \{\mathsf{Bet},\ \mathsf{Flip},\ \mathsf{Outcome},\ \mathsf{Win}\},\\
&R = \{\langle \mathsf{Bet}, \mathsf{Win}\rangle,\ \langle \mathsf{Bet}, \mathsf{Flip}\rangle,\ \langle \mathsf{Flip}, \mathsf{Outcome}\rangle,\ \langle \mathsf{Outcome}, \mathsf{Win}\rangle\}\\
&E = \{\mathsf{Flip} = 1 \Leftrightarrow \mathsf{Bet} \neq u,\\
&\quad \mathsf{Outcome} = u \Leftrightarrow \mathsf{Flip} = 0,\\
&\quad \mathsf{Flip} = 1 \Rightarrow \mathbf{p}(\mathsf{Outcome} = x) = 0.5, \text{ for } x \in \{1, 2\},[9]\\
&\quad \mathsf{Win} = 1 \Leftrightarrow \mathsf{Outcome} = \mathsf{Bet} \text{ and } \mathsf{Bet} \neq u\},\\
&\quad \mathsf{Win} = 0 \Leftrightarrow \mathsf{Outcome} \neq \mathsf{Bet} \text{ and } \mathsf{Bet} \neq u\}\\
&A = \{A(\mathsf{Bet}) = 1, A(\mathsf{Flip}) = 1, A(\mathsf{Outcome}) = 2, A(\mathsf{Win}) = 0\}
\end{aligned}
$$

[7] Further constraints on membership conditions for the law sets of these causal models are introduced via the Nomicality constraint offered in Sect. 3.

[8] Where I think it will not cause confusion, I let linguistic constructions act as names for themselves.

[9] Hiddleston directs us to read '⇒' as a strict conditional. (Cf. Hiddleston 2005, p. 651.) I see no reason not to read it as the material conditional.

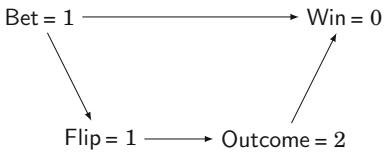



**Fig. 1** The causal model depicting the causal system described in *Betting on Coin Tosses*

A few further comments about $M$ are in order. Bet takes as values $u$ (no bet), 1 (tails), or 2 (heads). Outcome takes $u$ (no outcome), 1 (tails) or 2 (heads). Flip and Win each take 0, 1, or $u$ (No/Yes/Neither, respectively) as values. $M$ may be depicted pictorially as in Fig. 1. Taking $M$ to interpret the events described in *Betting on Coin Tosses*, $M$ contains variables set to their actual values.

Loosely put, if one switches the value of Bet from 1 to 2 while making as few further changes as possible to the values of the variables of $M$ in such a way as to preserve consistency with the laws of $E$, then the value of Win must be switched to 1. So TC tells us that the counterfactual Bet-Heads > Win is true, as it should. At this point, Hiddleston's theory deserves precise formulation. To this end, a number of definitions need stating[10]:

> **Parent variables** The set of parents $pa(X)$ of an event variable $X$ is the set containing (as members) all and only variables $W$ such that for some pair $\langle Y, X\rangle \in R$, $W = Y$.

Let $X$ be a parent of $Y$ in $M$, $X = x$, and $Y = y$. We denote $pa(Y) - \{X\}$ by $Z^{\rightarrow}$ and the values in $M$ of its members by $z^{\rightarrow}$. (So [given that $X$ is specified] $Z^{\rightarrow} = z^{\rightarrow}$ denotes a partial assignment of values to the variables in $M$—the assignment $A$ minus the ordered pair that gives the value of $X$ in $M$.) Then we may characterize *Direct positive influence* as follows.

> **Direct positive influence** (**formulation intended to capture contrastive facts**): $X = x$ *in comparison* to $X = x'$ has direct positive influence on $Y = y$ in $M$ iff (holding fixed $M$'s laws)
>
> $$\mathbf{p}(Y = y | X = x \ \&\ Z^{\rightarrow} = z^{\rightarrow}) > \mathbf{p}(Y = y | X = x' \ \&\ Z^{\rightarrow} = z^{\rightarrow}).$$
>
> Let the *positive parents* of $Y$ in $M$ be $ppa(Y)_{M,\, M_i} = \{X : A(X)$ in comparison to $A_i(X)$ has direct positive influence on $Y = y$ in $M\}$ (Hiddleston 651).
>
> **Descendant variables** Call the converse of the parent relation the *child* relation. Then the *descendant* relation is the transitive closure of the child relation.
>
> **Causal break** A *causal break* in model $M_i$ relative to $M$ is a variable $Y$ such that $A_i(Y) \neq A(Y)$, and for each $X \in ppa(Y)_{M,M_i}$, $A_i(X) = A(X)$. $Break(M_i, M) = \{Y : Y$ is a causal break in $M_i$ from $M\}$.
>
> **Intact** $Intact(M_i, M) = \{Y : A_i(Y) = A(Y)$ and for each $X \in ppa(Y)_{M,M_i}$, $A_i(X) = A(X)\}$

[10] The list of definitions is Hiddleston's, though my formulations vary from his in places.

Following Hiddleston, let us suppose throughout that the antecedents of interest are either atomics, negations, or conjunctions (Hiddleston 2005, p. 642).[11] According to TC, the truth conditions for a counterfactual $\varphi > \psi$ interpreted via a case $C$ are given in terms of a good model $M$ of $C$ and the minimal alterations of $M$ that satisfy $\varphi$. We say that a model $M_i$ is *minimally altered* (relative to $M$ and an antecedent $\varphi$) if $M_i$ meets the following conditions:

**$\varphi$-Minimal model:**

(i) $M_i$ satisfies the antecedent $\varphi$.

(ii) For the set of variables $Z$ comprised of variables $V \in \boldsymbol{V}$ such that $V$ is a non-descendant of all variables denoted by $\varphi$, $Intact(M_i, M) \cap Z$ is *maximal* among $\varphi$-models, where $Intact(M_i, M) \cap Z$ is maximal among $\varphi$-models iff no $\varphi$-model $M_k$ is such that $Intact(M_i, M) \cap Z \subset Intact(M_k, M) \cap Z$.

(iii) $Break(M_i, M)$ is *minimal* among $\varphi$-models, where $Break(M_i, M)$ is minimal among $\varphi$-models iff there is no model $M_k$ such that $Break(M_i, M) \supset Break(M_k, M)$.

All variables that are not break variables or intact are, using Hiddleston's terminology, said to be *up for grabs variables*.

Definitions at hand, we can now "officially" evaluate the counterfactual Bet-Heads > Win via TC. Enquiring after the set of $\mathsf{Bet} = 2$-minimal models (relative to $M$) shows that there is but one such model, $M_i$, (relative to $M$). $Break(M_i, M) = \{\mathsf{Bet}\}$, $Intact(M_i, M) = \{\mathsf{Flip}, \mathsf{Outcome}\}$, and Bet-Heads > Win (i.e., $\mathsf{Bet} = 2 > \mathsf{Win} = 1$) comes out true because $M_i$ (as depicted in Fig. 2) verifies $\mathsf{Win} = 1$.

An important point to note is that Hiddleston's semantics returns the right truth value for Bet-Heads > Win in substantial part because of Hiddleston's concept of direct positive influence and the role it serves in defining the concept of a minimal alteration. Notice that because the probability of heads given a coin flip equals the probability of tails given a coin flip, $\mathsf{Flip} = 1$ is not a positive parent of $\mathsf{Outcome} = 2$ (in $M'$ relative to $M$). Hence, Outcome has no positive parents and a relevant alteration of $M$ such that $\mathsf{Outcome} \neq 2$ will contain a larger break set than $M'$. These considerations serve to illustrate the important role of laws describing probabilistic dependency relations and the definition of 'positive parent' in Hiddleston's theory. (It is the definition of 'positive parent' above that enables Hiddleston's theory to handle both Morgenbesser and non-Morganbesser cases alike and, thus, to avoid the problem that Lewis (1979a) highlights in his parenthetical remarks on his page 472 and that led Lewis to include the unsatisfying (4) among his weighted constraints.)

Keeping the above in mind, in evaluating the cases I utilize below we can, for convenience, utilize the following simpler but approximate characterization (also owed to Hiddleston) of direct positive influence:

> **Direct positive influence** (**simpler but approximate formulation**): $X = x$ has direct positive influence on $Y = y$ in $M$ iff (holding fixed $M$'s laws)]

[11] Here I reproduce Hiddleston's endnote 9: "Disjunctive antecedents pose problems for all theories of counterfactuals and I do not have anything to add concerning them. See, e.g., Loewer (1976)".

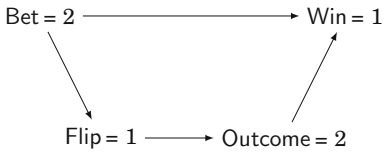



**Fig. 2** The single $\varphi$-minimal model, $M_i$, relevant to evaluating the counterfactual Bet-Heads > Win. $Break(M, M_i) = \{\mathsf{Bet}\}$ $Intact(M, M_i = \{\mathsf{Flip}, \mathsf{Outcome}\})$. And $A_i(\mathsf{Win}) = 1$ So $v$([Bet-Heads > Win]) $= True$

$$\mathbf{p}(Y = y|X = x \;\&\; Z^{\rightarrow} = z^{\rightarrow}) > \mathbf{p}(Y = y|X \neq x \;\&\; Z^{\rightarrow} = z^{\rightarrow}).$$

An attractive feature of Hiddleston's account is that it handles backtrackers very well. It is easy to verify, for instance, that TC returns the value True for the counterfactual "If the coin had come up tails but Ben had lost [$\mathsf{Outcome} = 1$ & $\mathsf{Win} = 0$] then it would have been that he bet heads [$\mathsf{Bet} = 2$]". Another feature of the account is that any counterfactual containing an antecedent inconsistent with the law set $E$ comes out vacuously true.[12] This is because, in such cases, the set of $\varphi$-minimal models is empty, so any consequent is verified by all the members of that set (vacuously).

### 3.1 Counterlegals and TC

Counterlegals are counterfactuals containing an antecedent that is inconsistent with some set of laws. Counterlegals often figure in non-trivial reasoning; (c) is a case in point.

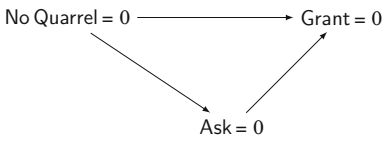
(c) If those neutrinos had travelled faster than the speed of light, then relativistic physics would be inaccurate.

Thinking back to 2011 and its media reports of (now known to be spurious) evidence that some neutrinos had exceeded the speed of light, it is easy to find oneself considering what might have been, had we come to know that some neutrinos travel faster than light. To do so is to engage in non-trivial counterlegal reasoning (assuming the relevant laws are as we believe them to be). This is, I submit, strong prima facie motivation for desiring a semantics that facilitates non-trivial evaluation of counterlegals.

Lewis' theory has the attractive feature that evaluating counterlegals often poses no special difficulty. To evaluate a counterlegal on the Lewisian picture (and accepting the limit assumption for a moment), simply query the nearest nomically-impossible worlds. But, as noted already, Hiddleston's theory counts any counterfactual that is counterlegal (relative to a given model $M$) as vacuously true (relative to $M$). This is just because for a counterfactual to be counterlegal relative to $M$ is for it to have an antecedent inconsistent with the laws of $M$. This, of course, means that all the models having the same laws as $M$ that satisfy the counterlegal antecedent (of which there are none) satisfy the consequent, for any consequent at all.

[12] Hiddleston points this out in his endnote 7 (Hiddleston 2005, p. 655).

**Fig. 3** $M_Q$




Here I present a now well-known case of Downing's (1958–1959) and argue that treating all counterlegals as true is a significant limitation of TC. Downing utilized his case in an effort to show that combining backtracking and forward-tracking reasoning in the same context is a recipe for inconsistency. I utilize the case to make a point about counterlegals—they deserve non-trivial analysis–and a point about TC—it leaves something to be desired with respect to counterlegals. Here is Bennett's (2003) Austenian modification of the case:

> ***Mr. D'Arcy and Elizabeth*** Mr D'Arcy and Elizabeth quarrelled yesterday, and she remained angry with him this morning. So *If he had asked her for a favour this morning, she would have refused it*. On the other hand, he is a proud man …[H]e would never risk being turned down; so if he had asked her for a favour this morning, they wouldn't have quarrelled yesterday, in which case Elizabeth would have been her usual accommodating self and would have granted the favour. So *If he had asked her for a favour this morning, she would have granted it* (Bennett 2003, p. 205).[13]

The story seems to equally well support both $\mathsf{Ask} = 1 > \mathsf{Grant} = 1$ and $\mathsf{Ask} = 1 > \mathsf{Grant} = 0$. But these two counterfactuals cannot be true together, so if a semantics decides in favor of one of them, it should decide against the other. TC legislates in favor of $\mathsf{Ask} = 1 > \mathsf{Grant} = 1$ and against its competitor. This is because, in the presence of the law set recoverable from the story, there is no admissible assignment that sets $\mathsf{Ask}$ to 1 and $\mathsf{No\ Quarrel}$ to 0. Lewis' theory, on the other hand, delivers the opposite result because the argument to the truth of $\mathsf{Ask} = 1 > \mathsf{Grant} = 0$ rests on appeal to the backtracker: "If he had asked her for a favour this morning, they wouldn't have quarrelled".

We can interpret counterfactuals concerning the story using a model $M_Q$, as depicted in Fig. 3. The law set for $M_Q$ is: $E_Q = \{$(a) $\mathsf{Grant} = \mathsf{No\ Quarrel} \times \mathsf{Ask}$, (b) $\mathsf{No\ Quarrel} = 0 \Rightarrow \mathsf{Ask} = 0\}$

Here is a problem for TC: There is a perfectly semantically-respectable pair of counterfactuals which we can extract from the story that are contraries of each other, neither of which TC will allow us to evaluate non-trivially. Consider the counterfactuals: "If there had been a quarrel and Mr. D'Arcy were to ask Elizabeth for a favor, then she would not grant it" and "If there had been a quarrel and Mr. D'Arcy were to ask Elizabeth for a favor, then she would grant it". The first seems true, but non-trivially and the second seems false. TC decrees that both are (vacuously) true. So, TC appears

[13] I believe it is correct to think that the context, and hence the operative laws, change as the *Mr. D'Arcy and Elizabeth* story plays out. The points I make are intended to stand relative to the single context we are left with at the close of the story.

to let us down in two ways; it counts a false counterfactual as true, and it evaluates trivially some counterfactuals that it should not.

Of course, if it's a live possibility that Mr. D'Arcy and Elizabeth have a fight and Mr. D'Arcy still asks a favor, then $M$ is not a good model for the case, since $M$'s laws rule out such a state. Fair enough. But, let $M$ in fact be a good model for the case. (It's fiction after all.) My present point is simply that, even still, the counterfactuals corresponding to the question "What would happen if they had a fight and Mr. D'Arcy asked for a favor?" remains perfectly coherent *semantically* and non-trivial, but Hiddleston's theory is unable to accommodate it as such.

Moreover, even when the law-like generalities at issue come to us not from some fictional story but, rather, from our most sober pictures of the world, we still meet counterfactuals properly evaluated counterlegally relative to such pictures. Take, for example, a variation of the *Tracking Photons* case: Suppose we are physicists at work in our lab and that we've just detected the arrival of a photon at point $B$ at time $t_3$. I say "If that photon had arrived a moment earlier, then our careers would have been made". If you backtrack as required by TC and reply that I am mistaken because, "Had that photon arrived any earlier it would have been fired earlier" you've simply misinterpreted my remark. Further still, the intended counterlegal reading can easily be made explicit, for I could have simply said "If that photon had been fired at the time and place it actually was and had arrived where it did a moment earlier, then our careers would have been made". But then TC tells us inappropriately that the aforementioned counterfactual and the counterfactual resulting from negating its consequent are both true.

The *Mr. D'arcy* case and the above variation on *Tracking Photons* also bring to the fore issues about pragmatic effects and counterlegals. I believe it is probably correct that accomodation effects [in Lewis' (1979b) sense] are often such that an utterance of a counterfactual will change the context so that it can be interpreted non-counterlegally. This is the effect that I sought to guard against when I stipulated that $M$ above counts as a good (i.e., accurate) model for the *Mr. D'arcy* case. But we don't always accommodate so as to facilitate a non-counterlegal reading. The variation on *Tracking Photons* may be taken to show that. It is an interesting and complex question—in pragmatics—under which circumstances we read a counterfactual counterlegally and under which cases we accommodate to avoid doing so. But I take this question to lie outside the scope of the semantic project of this paper, consequently I will ignore or fix for these issues to the extent that I am able. One device for encouraging the counterlegal reading I am after is to parenthetically insert the word 'miraculously' into a counterfactual—as in "If that photon had (miraculously) arrived a moment earlier, then our careers would have been made". This is a device I will employ below.

## 4 Evaluating counterlegals

How might we develop a causal-model semantics for both backtracking and counterlegal counterfactuals? Here are some suggestive remarks. Note first off that, as far as the formal semantics is concerned, legality is relative to a model's graph and law set—following Hiddleston, call such a pair a schema (Hiddleston 2005, p. 639). So we

can say that legality is schema-relative. I think some proposal along the following lines will work: To evaluate a counterlegal we first locate the class of schemas, $\mathscr{S}_{illegal}$, that are consistent with the antecedent yet minimally unlike the schema $S = \langle G, E\rangle$ contained in the original model $M$. We then consider the class of models built on the members of $\mathscr{S}_{illegal}$ by adding assignments that are minimally unlike the assignment $A$ contained in $M$. We might then say that a counterlegal $\iota > \psi$ is true just in case $\psi$ is true in all minimally-illegal models (relative to $M$). The challenge, of course, is to define this notion of minimal illegality in such a way that it systematically returns the right truth values while, ideally, staying true to some intuitive picture of what one plausibly does when considering a counterlegal.

Here is a (loose and informal) first crack at describing a procedure for locating the class of minimally-illegal models:

> ***First Crack*** To evaluate a counterfactual containing an antecedent inconsistent with the laws of the causal model that interprets it, subtract as few members of the law set as possible in order to obtain a law set consistent with the antecedent. Then, if all such models that can be "built on" such schemas and that are otherwise minimally altered satisfy the consequent, call the counterlegal true.

*First Crack* works for some cases, but falls to counterexamples. Consider the following case.

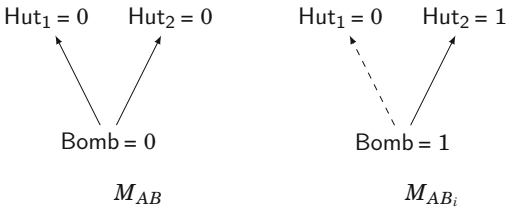
> ***Atomic Bombs*** Two straw huts, $\text{Hut}_1$ and $\text{Hut}_2$, stand on either end of an island. Intermediate between the huts is a very large hydrogen bomb, B. It is nomically impossible for either $\text{Hut}_1$ or $\text{Hut}_2$ to withstand B's blast.

For all that has been said so far, a good law set for the Atomic Bombs case is $E_{AB}$. (Let the value 1 encode exploding.)

$$E_{AB} = \{(a)\mathsf{B} = 1 \Rightarrow \mathsf{H}_1 = 1, (b)\mathsf{B} = 1 \Rightarrow \mathsf{H}_2 = 1\}$$

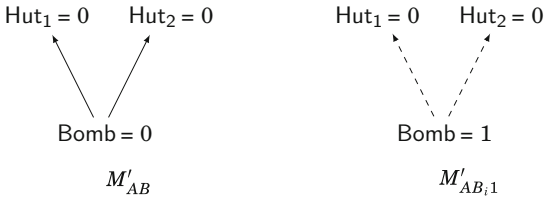
We can model *Atomic Bombs* using $M_{AB}$ of Fig. 4. Now consider the counterlegal, "If B were to explode but Hut 1 were to (miraculously) remain standing, then Hut 2 would (still) be destroyed" $[\mathsf{Bomb} = 1\ \&\ \mathsf{H}_1 = 0 > \mathsf{H}_2 = 1]$. If I may adopt a worlds picture temporarily, to assess this counterlegal, we should imagine a case satisfying the antecedent at those worlds most similiar to ours, similarity of causal-dependency relations included and important. Hence, if it is really a law that whenever a bomb like B goes off, huts like 1 and 2 are destroyed, then the relevant illegal case should be one in which the above counterlegal comes out true.

I am not entirely convinced by the above reasoning and think that more should be said. Why not think that if Hut 1 were to withstand the blast, then Hut 2 might also withstand it, and so the conditional at issue is false? Indeed, I think there may be a reading, relative to some contexts, of $\mathsf{B} = 1\ \&\ \mathsf{H}_1 = 0 > \mathsf{H}_2 = 1$ such that it comes out false. But this is not, I suggest, a counterlegal reading. On such a non-counterlegal reading the model used to interpret the conditional would need to encode the possibility of the bomb going off and Hut 2 withstanding the blast. On the other hand, relative to contexts where $M_{AB}$ is a good model for the case and $\mathsf{B} = 1\ \&\ \mathsf{H}_1 = 0 > \mathsf{H}_2 = 1$ is interpreted relative to $M_{AB}$, then I think $\mathsf{B} = 1\ \&\ \mathsf{H}_1 = 0 > \mathsf{H}_2 = 1$ should come out true. In any case, this is the hypothesis I will explore.




$M_{AB}$ $M_{AB_i}$

**Fig. 4** The models $M_{AB}$ and $M_{AB_i}$, with law sets as follows. $E_{AB} = \{(a)\ B = 1 \Rightarrow H_1 = 1, (b)\ B = 1 \Rightarrow H_2 = 1\}$ $E_{AB_i} = \{(b)\ B = 1 \Rightarrow H_2 = 1\}$




$M'_{AB}$ $M'_{AB_i1}$

**Fig. 5** The models $M'_{AB}$ and $M'_{AB_i1}$, with law sets as follows. $E'_{AB} = \{(\alpha)\ B = 1 \Rightarrow H_1 = 1 \wedge H_2\} = 1$, $E'_{AB_i1} = \emptyset$

First Crack at first appears to give the desired result, as illustrated in $M_{AB_i}$ of Fig. 4. I will use dotted lines in place of arrows to depict dependency relations that are no longer characterized by any law sentence. With respect to the causal models, whether such relations should be thought of as having been removed or rather as simply having become inert is a point of indifference I think. For simplicity, in the developments to follow I will treat them as present but inert.

*First Crack* fails, but the way it fails is revealing: Consider $M'_{AB}$; this model is identical to $M_{AB}$ save for its law set $E'$. Importantly, however, $E'$ is logically equivalent to $E$ (as may be verified by comparing the law set corresponding to $M'_{AB}$ of Fig. 5 with that of $M_{AB}$ depicted in Fig. 4). But, applying *First Crack* to $M'_{AB}$ yields the value False for $\mathsf{B} = 1\ \&\ \mathsf{Hut_1} = 0 > \mathsf{Hut_2} = 1$. This is because relative to $M'_{AB_i1}$, applying *First Crack* in service of an evaluation of the counterlegal at issue leads us to look to $M'_{AB_i1}$, and in $M'_{AB_i1}$ the value of $\mathsf{Hut_2}$ is entirely independent of the variables implicated in the antecedent at issue. Since the law sets of $M_{AB}$ and $M'_{AB}$ are equivalent in a rather strong way and *First Crack* gives no way of choosing between them, we have a refutation of *First Crack*.

First Crack seems to let us down because it allows modifications to law sets that do not appropriately mirror the structure of the dependency relations depicted in the causal model at issue. One strategy for ameliorating this problem while adhering to the subtract-sentences-first style approach of *First Crack*, involves restricting the kinds of sentences that a law set may contain so that the admissible law sets must contain laws that reflect something of the structure of the dependency relations depicted in the models. The thought, then, would be that models like $M'_{AB}$ could then be made inadmissible. This is a strategy which I will say more about below. Another idea worth exploring is that of beginning to evaluate counterlegals by subtracting as few arrows as possible, rather than first focusing on modifying the law set. We might motivate a strategy of the arrow-subtraction-first sort by suggesting, as above, that what went wrong with first crack is that it directs us to modify the sentential structure of the law set without regard to the structure of the dependency relations given by the model. Here is a procedure along these latter lines (merely described informally for the moment):

***Second Crack*** To evaluate a counterfactual containing an antecedent inconsistent with the laws of the causal model that interprets it, subtract as few arrows as possible in order to obtain a structure that doesn't rule out the antecedent. Then construct a law set preserving "as many" of $E$'s original entailments as possible. If all such resulting models that are otherwise minimally altered satisfy the consequent, call the counterlegal true.

The characterization of causal models given so far forces no changes to be made to the members of $E$ given any subtraction of some arrows. But for *Second Crack* to work, some condition is needed that constrains the admissible laws in relation to the members of $R$ (the arrows). Here is such a constraint:

***Nomicality**** $\forall\varphi \in E$, $\varphi$ is of the form $\chi[V_n = v_n, \ldots, V_k = v_k] \Rightarrow \psi[\mathbf{p}(V_i = v_i) = r_i, \ldots, \mathbf{p}(V_j = v_j) = r_j]$ containing exactly one occurrence of $\Rightarrow$,[14] and every variable denoted by $\varphi$'s antecedent is either a parent or child of some variable denoted by the consequent and every variable denoted by the consequent is a parent or child of some variable denoted by the antecedent. (Let law sentence containing '$\Leftrightarrow$' or '$\times$' or containing no probability operators abbreviate the obvious law sentences having the appropriate truth conditions and satisfying *Nomicality**.)

*Nomicality** is meant to guarantee that any sentence in a law set $E$ directly characterizes some dependency relations depicted in the model, rather than a "mere entailment" of some such relation(s). Supplemented with *Nomicality**, *Second Crack* has us look first to the depicted nomic-dependency relations in our best picture of a part of the world (precisified as a causal model) Next, we make minimal changes to this picture as required. Last we recover as much of the original linguistic characterization of the depicted dependency relations as possible. That is, we try to recover a law set that preserves an inclusion-maximal subset of the entailments of the initial law set.

But once *Nomicality** is enforced, *Second Crack* becomes too ham fisted. If we care to interpret embedded conditionals and we care to interpret backtrackers, with *Nomicality** in place we will confront cases where Second Crack counsels us to remove an arrow between two variables, say $\mathsf{A}$ and $\mathsf{B}$, where $\mathsf{A} = 1$ is causally necessary and sufficient for $\mathsf{B} = 1$. *Nomicality** will then force us to lose the law sentences corresponding to both conditional necessity and conditional sufficiency for $\mathsf{A} = 1$ relative to $\mathsf{B} = 1$, and in some cases this is to lose too much. Here is such a case:

***Hiddleston's Cannon*** Consider a ceremonial cannon. Let background conditions be such that, while the fuse is merely a contributing cause of cannon explosions, cannon explosions invariably cause the occurrence of a flash and a bang and the only cause of a cannon flash or bang is a cannon explosion (Fig. 6).[15]

[14] Where $\Rightarrow$ is the material conditional of our metalanguage.

[15] The imperfectness of the fuse as a cause of explosions was important for Hiddleston but is unimportant for my purposes. I have simply chosen to leave this feature of the case unchanged. This is why the probability statement appears in the consequent of ($\alpha$); it may be ignored.

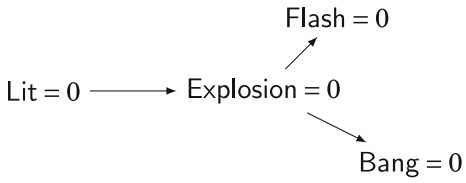
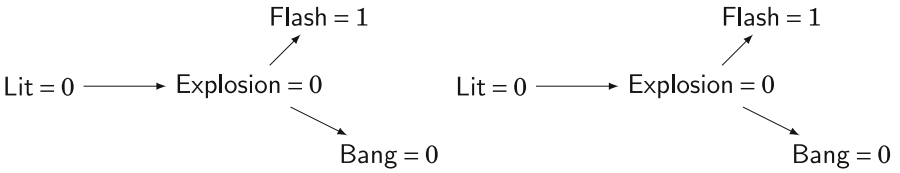



**Fig. 6** $M_{HC}$. $E_{HC} = \{(\alpha)\ \mathsf{L} = 1 \Rightarrow \mathsf{p}(\mathsf{E} = 1) = .95, (\beta)\ \mathsf{E} = 1 \Rightarrow \mathsf{F} = 1 \wedge \mathsf{B} = 1, (\gamma)\ \mathsf{F} = 1 \vee \mathsf{B} = 1 \Rightarrow \mathsf{E} = 1\}$




$M_{HCi1}$ $M_{HCi2}$

**Fig. 7** The two $\mathsf{F} = 1 \wedge \mathsf{B} = 0$-minimally illegal models $M_{HCi1}$ (relative to $M_{HC}$ above) and $M_{HCi2}$, with corresponding law sets below. Both models satisfy $(\mathsf{B} = 1 > \mathsf{E} = 1)$. Hence $M_{HC}$ satisfies $(\mathsf{F} = 1 \wedge \mathsf{B} = 0) > (\mathsf{B} = 1 > \mathsf{E} = 1)$. $E_{HCi1} = \{(\alpha)\ \mathsf{L} = 1 \Rightarrow \mathsf{p}(\mathsf{E} = 1) = .95, (\beta)\ \mathsf{E} = 1 \Rightarrow \mathsf{F} = 1 \wedge \mathsf{B} = 1, (\gamma')\ \mathsf{B} = 1 \Rightarrow \mathsf{E} = 1\}$, $E'_{HCi2} = \{(\alpha)\ \mathsf{L} = 1 \Rightarrow \mathsf{p}(\mathsf{E} = 1) = .95, (\beta')\ \mathsf{E} = 1 \Rightarrow \mathsf{F} = 1, (\gamma)\ \mathsf{F} = 1 \vee \mathsf{B} = 1 \Rightarrow \mathsf{E} = 1\}$

Consider now the counterlegal: "If the cannon flashed but made no bang, then if it had made a bang there would have been an explosion".[16] I claim that (after some reflection perhaps) this counterlegal (the consequent of which is a backtracker) should be judged true. This is because it seems that the relevant—loosely, the nearest—nomically-impossible scenarios involving a cannon flash but no cannon bang are either states in which a cannon flash is not necessary for a cannon explosion or they are states such that cannon explosions are not sufficient for cannon bangs. Further, I see no reason for judging one of the disjuncts as "less illegal". The crucial point though is that neither of these allowances suggests that cannon explosions would no longer be necessary for bangs. So we should judge that $(\mathsf{F} = 1\ \&\ \mathsf{B} = 0) > (\mathsf{B} = 0 > \mathsf{E} = 1)$ is true. But that means the arrow-subtraction-first approach of *Second Crack* returns the wrong evaluation. This is because according to *Second Crack* all $\mathsf{F} = 1\ \&\ \mathsf{B} = 0$-minimally illegal models (relative to $M_{HC}$) have an arrow removed, either between Explosion and Flash or between Explosion and Bang. So, by *Nomicality**, there is a minimally illegal model depicting a bang without an explosion (Fig. 7).

What of *Nomicality**? It too can't be quite what is wanted. This is because there is an easy trick for constructing sentences that, intuitively, do not appropriately characterize the edges of a causal model yet nevertheless satisfy *Nomicality**. Consider a simple graph containing three binary variables A, B, C, an arrow from A to B, but no arrows into or out of C. Then the sentence $\mathsf{A} = 1 \Rightarrow \mathsf{C} = 1$ is not admissible by *Nomicality**. However, the sentence $\mathsf{A} = 1 \Rightarrow [\mathsf{C} = 1\ \&\ (\mathsf{B} = 1 \vee \mathsf{B} = 0)]$ is admis-

[16] This is a right-nested conditional, and Hiddleston (2005) does not consider such conditionals. In considering nested conditionals we may be going beyond the class of conditionals Hiddleston had in mind to treat with his semantics.

sible and effectively equivalent to the previous sentence. So we need a more restrictive constraint.[17] I propose the following:

> ***Nomicality*** $\forall\varphi \in E$, $\varphi$ is of the form $\chi[V_n = v_n, \ldots, V_k = v_k] \Rightarrow \psi[\mathbf{p}(V_i = v_i) = r_i, \ldots, \mathbf{p}(V_j = v_j) = r_j]$ containing exactly one occurrence of $\Rightarrow$ and every variable denoted by $\varphi$'s antecedent is a parent of every variable denoted by $\varphi$'s consequent, or every variable denoted by $\varphi$'s consequent is a parent of every variable denoted by $\varphi$'s antecedent. (Let law sentences containing '$\Leftrightarrow$' or '$\times$' or containing no probability operators abbreviate the obvious law sentences having the appropriate truth conditions and satisfying *Nomicality*.)

Taking stock of where things stand now: This last formulation of the *Nomicality* constraint seems to solve the problem just described; and the arrow-subtraction-first approach of *Second Crack* has been rejected. We can return to a sentence-subtraction-first approach, but this is because *Nomicality* appears sufficiently restrictive to guarantee that subtracting minimal subsets of law sentences will allow sufficiently fine-grained manipulation of the dependency relations depicted by the causal model at issue. That is, the lesson of *First Crack's* failure is still respected—how we modify law sets under counterlegal reasoning must be sensitive to the structure of the dependency relations depicted by the edges of the causal-model we use to represent the case at issue. But so too, the lesson of *Second Crack's* failure should be respected—changing our linguistic characterizations of dependency relations allows for finer manipulations of dependency-relation structures than simply ignoring such relations entirely.[18]

What follows now is the official proposal. The proposal takes the form of a definition of minimal illegality followed by a specification of truth conditions for counterlegals in a causal-model semantics in terms of minimal illegality. Holding *Nomicality* fixed (i.e., with variables restricted to causal models satisfying *Nomicality*), we have:

> ***Minimal Illegality*** A model $M_i = \langle S_i = \langle G_i, E_i\rangle, A_i\rangle$ is minimally illegal relative to $M = \langle S = \langle G, E\rangle, A\rangle$ and a counterlegal $\iota > \psi$ iff:[19]
>
> 1. $M_i$ is an $\iota$-model (i.e., $M_i$ satisfies $\iota$);
> 2. $G_i = G$ and $ENT(E_i) \subseteq ENT(E)$ (where '$ENT(E)$' denotes the classical closure of the law set $E$.);
> 3. but $\neg\exists M_k$: $M_k$ satisfies conditions 1. and 2. and $ENT(E_k) \supset ENT(E_i)$;
> 4. and $\forall M'$ s.t. $M'$ satisfies $\iota$ and $S' = S_i$ : $\neg[Intact(M_i, M) \cap Z \subset Intact(M', M) \cap Z]$ and $\neg[Break(M_i, M) \supset Break(M', M)]$ (i.e., $M'$ is *otherwise minimally altered*).

Conditions 1.–3. describe what it takes for a model to be appropriately illegal in order to be a candidate for minimal illegality. Intuitively, they tell us what it is for a model to violate as few laws as possible while satisfying the antecedent of interest.

Condition 1. tells us that $M_i$ must satisfy the antecedent. Condition 2. tells us that it's graph must be the same as $M$'s, and the (classical) closure of its law set must

[17] Thanks to Adam Edwards and Adam Sennet for making me think more carefully about these points.

[18] This is not to say that we must work with linguistic characterizations of the relations, we could consider changes to the relations directly.

[19] Recall that for a model $M$ we call the pair containing its graph and law set a schema and denote it with an '$S$'.

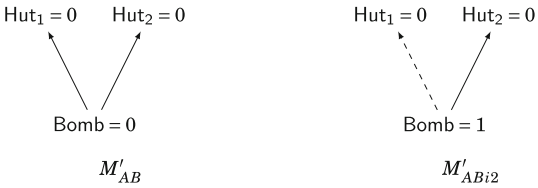



**Fig. 8** The models $M'_{AB}$ and $M'_{ABi2}$, with corresponding law sets. $E'_{AB} = \{(\alpha)$ $B = 1 \Rightarrow H_1 = 1 \wedge H_2 = 1$, $E'_{ABi2} = \{(\alpha'),\ B = 1 \Rightarrow$ $H_2 = 1\}$

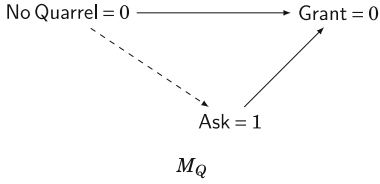



**Fig. 9** $E_Q = \{$(a) Grant = No Quarrel × Ask$\}$

be a subset of the closure of $M$'s law set. Condition 3. tells us that among models satisfying the first two conditions, the closure of $M_i$'s law set is inclusion maximal relative to that of $M$. Condition 4. tells us what it is for an appropriately illegal model to be otherwise minimally altered. Recall that $Z$ is the set of variables that are non-descendants of the variables "mentioned" by the antecedent. So, 4. tells us what it is for an appropriately-illegal model to maximize its set of relevant intact variables and minimize its break variables among appropriately-illegal models.

> ***Counterlegal truth conditions*** A counterlegal $\iota > \psi$ is true iff every minimally illegal model that satisfies $\iota$ also satisfies $\psi$.

What then of the earlier problem cases, *Mr. D'arcy and Elizabeth*, *Atomic Bombs*, and *Hiddleston's Cannon*? With *Nomicality* in place, the proposed semantics for counterlegals, CTC[20], appropriately handles the *Atomic Bombs* case that proved problematic for first crack. To appreciate this, notice that according to CTC, relative to $M'_{AB}$ (as depicted in Figs. 5 and 8), all $\mathsf{Bomb} = 1$ & $\mathsf{Hut}_1 = 0$-minimally-illegal models will satisfy $\mathsf{Hut}_2 = 1$. This is because CTC counsels us to subtract a sentence that then makes the arrow inert between $\mathsf{Bomb}$ and $\mathsf{Hut}_1$. Moreover, Nomicality rules $(\alpha)$ inadmissible, but $(\alpha')$ remains an admissible entailment of the original set $E'_{AB}$, so no model not entailing $(\alpha')$ will count as minimally illegal. So CTC gets *Atomic Bombs* right, yielding $v(\mathsf{B}_2 = 1\ \&\ \mathsf{H}_1 = 0 > \mathsf{H}_2 = 1) =$ True, even when interpreted via $M'_{AB}$. The embedded conditional of *Hiddleston's Cannon* also receives an appropriate evaluation. Returning to *Mr. D'Arcy and Elizabeth*, TC plus CTC correctly returns $v(\mathsf{No\ Quarrel} = 0\ \&\ \mathsf{Ask} = 1 > \mathsf{Grant} = 0) =$ True (Cf. Fig. 9).

To be explicit now about the manner in which backtrackers are treated in CTC, given a causal-model semantics like Hiddleston's TC, CTC functions to extend it as follows: *Minimal Illegality* is partially defined in terms of the given criteria for $\varphi$-minimality. This is done so that non-counterlegals get evaluated simply as degenerate cases of counterlegals. That is, they are simply cases such that subtracting the empty set of laws suffices to secure legality. For such non-counterlegal cases, the new semantics,

[20] CTC for *counterlegal truth conditions* or *causal theory of counterlegals*, depending on your taste.

CTC, simply reduces to TC. As long as TC handles non-counterlegal backtrackers, so too does CTC.

## 5 CTC, TC, and interventionism

Before concluding I want to draw a contrast between CTC, Hiddleston's TC, and Pearl-style (2000) interventionist semantic theories for counterfactuals. To this end, I will say a little about Pearl-style approaches and then consider a final set of counterfactuals relative to a final case.

Pearl's (2000) causal-model semantics is perhaps the most well-known and influential causal-model semantics for subjunctive conditionals. Following Briggs (2012), I will call Pearl-style semantic theories of counterfactuals, *interventionist* theories. Informally put, on an interventionist approach, one evaluates a counterfactual $\varphi > \psi$ relative to a model $M$ by changing the values of some variables in $M$ in such a way that the model satisfies $\varphi$, while holding fixed the values of all non-descendant variables not denoted by $\varphi$. Next one lets these changes flow causally downstream allowing them to effect changes to the values of descendant variables in the manner required to preserve consistency with the unaltered equations (roughly the laws) of $M$.[21] Such an evaluation procedure can be though of as simulating a change to a causal system at the hands of an external force—the idea of an experimental scientist manipulating a causal system is suggestive here. Indeed such a procedure is closely tied to how we investigate causal hypotheses.[22] Even with such a cursory characterization, it is not too hard to appreciate one way that interventionist evaluation procedures differ from that of Hiddleston's TC—on an interventionist approach any changes to upstream variables are precluded and, consequently, backtracking is never sanctioned.

Here is a case intended to draw out a contrast between Pearl-style approaches, Hiddleston's TC, and the semantics CTC developed in this paper.

> ***Flashlight*** Imagine I hold out a flashlight containing fresh batteries and in good working order. I do not flip the switch and the light does not turn on.

I believe all of TC, CTC, and Briggs' (2012) interventionist Semantics will count $M_F$ (Fig. 10) as an appropriate model for this case. Consider these three counterfactuals:

(1) If the switch had been flipped and the light had come on, then if I had removed the batteries the light would have come on.
(2) If the switch had been flipped and the light had come on and I had removed the batteries, the light would have come on
(3) If the switch had been flipped and the light had come on and I had removed the batteries, the light would not have come on.

[21] For careful characterizations of interventionistic semantic theories, see, e.g., Pearl (2000; esp pp. 33–38, 202–215) and Briggs (2012). See, e.g., Rips (2009) for a lengthier informal characterization of intervention semantic theories, there called "pruning theory". For a largely critical discussion of interventionist theories see Fisher (2016).

[22] Notice that even if it is correct to say that causal relations determine which counterfactuals are true and which are false, it does not follow that our best (epistemic) methods for discovering such relations will also turn out to be our best (semantic) methods for evaluating counterfactuals.

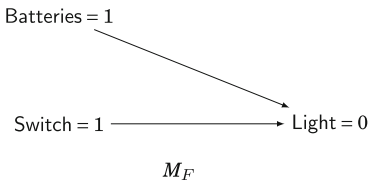



**Fig. 10** $E_F = \{\text{Light} = \text{Switch} \times \text{Batteries}\}$

(1) is false—if the switch had been flipped and the light had come on, it would nevertheless remain impossible for it to come on had I also removed the batteries. (2) has an antecedent that (relevantly) entails its consequent, so it is true. And (3) is false.

(1)–(3) are all right-embedded conditionals, but neither Pearl nor Hiddleston apply their semantic theories to such conditionals. However, as pointed out by Briggs (2012), a method for evaluating such conditionals is already implicit in most causal-model approaches: to evaluate a right-embedded conditional, locate a good model for the case, make that model satisfy the antecedent per the strictures of the semantic theory being applied, and then check whether the modified causal model would verify the right-embedded conditional per the strictures of the theory upon iteration of the evaluation procedure (Cf. Briggs 2012, p. 150). Indeed, among Briggs' (2012) primary tasks is that of rigorously laying out an interventionist semantics explicitly designed to interpret right-embedded counterfactuals.

Hiddleston's TC gets (1) and (2) correct, but it gets (3) wrong because TC trivially returns the value True for all counterlegals. Briggs' interventionist semantics gets (2) and (3) correct, but mis-evaluates (1) as true. Indeed *any* Pearl-style semantics applied to (1) will count it as true. This is because on an interventionist approach, intervening to set the value of the Light variable to 1 makes this variable-value pair independent of any of the event-type variables that it would otherwise be sensitive to. Consequently, an interventionist semantics will have no model that makes a sentence like (1) false. But the semantics CTC returns what I believe to be the desired values for all of (1)–(3).[23] To appreciate that CTC returns the value True for (1), notice that there is no way to make a causal model satisfy (1)'s antecedent without making the model satisfy (1)'s consequent. To appreciate that CTC gets things right for (2) notice that (put informally) to evaluate (2) we first look to $M_F$ and set the values of Switch and Light to 1. Call the resulting model $M_{F_{Switch,Batt}}$ Next we evaluate the right-embedded conditional relative to $M_{F_{Switch,Batt}}$ by "flipping" the value of Batteries to 0. But then to maintain consistency with the law set, Light must take the value 0. To appreciate that CTC gets the correct answer for (3), notice that it is a counterlegal, but no matter how the laws are broken there is no way to make a model satisfy (3)'s antecedent without falsifying its consequent.

[23] Note too that a typical Lewis-style possible-worlds semantics will agree with CTC on (1)–(3). This is enough, in the present context I hope, to calm suspicions that (1) is a backtracker or somehow unfairly non-standard as counterfactuals go. For further argumentation that counterfactuals like (1) are not backtrackers and often come out intuitively false, contrary to what interventionist semantic theories would predict, see Fisher (2016).

## 6 Conclusion

I've given some reasons to be dissatisfied with theories of counterfactuals that disallow non-trivial evaluation of backtrackers and counterlegals, and I've suggested a method for non-trivially evaluating counterlegals using a causal-model semantics. The proffered semantics appears to handle both backtrackers and counterlegals.

I close by considering some remarks of Bennett's (2003) to do with counterlegals and the prospects for developing a semantics for them. Bennett writes:

> If we are to have a single unified analysis that provides for counterlegals as well as counterfactuals…[we] need to find the closest A-worlds among the causally impossible ones; and we might do that on the basis of likeness to $\alpha$ in nomological structure.
> I would go along with this if I had a workable concept of nomological similarity; but I have not, and it seems that nobody else has either.
> The only philosopher I know of who has worked in detail on counterlegals is Pollock (1976, pp. 56–57, 93–97)…In giving details, he speaks of 'making deletions' in 'the actual set of basic laws', taking a minimal change to be one that deletes as few of the set's members as possible. This sounds all right until we ask: how can we count basic laws?…If we can solve the problem, it must be through the physicists' telling us things about the nomological structure of the actual world, as distinct from the propositional structure of their favourite theory. Objection: 'If it takes this kind of information to legitimize counterlegals, that puts them beyond the reach of most of us.' So they are, and so they should be (Bennett 2003, pp. 227–228).

Bennett suggests that in order to evaluate counterlegals we would need a way of assessing nomological similarity; he suggests that looking first to the set of laws partially characterizing a world and then making "minimal deletions" is not the right strategy to adopt because its success requires solving the apparently unsolveable task of sorting law sentences in order of basicness; and he suggests that an appropriate solution would have us look first to the nomological structure of the world (or at least our best picture of it) rather than the "propositional structure" of some favored theory. I agree with these points, and I believe they are lessons suggested by the failures of *First Crack* and *Second Crack*.

I part ways with Bennett when it comes to my optimism regarding the project of developing a unified semantics for forward-tracking counterfactuals, backtrackers, and counterlegals. This is in large part because I find it plausible that for all of us, not just the physicists among us, our counterfactual talk is guided by and interpreted relative to some favored pictures of the world, where these pictures include information about what may be thought of as nomological structures[24] The accuracy of such pictures is a separate issue of course.

In this paper I've highlighted and sought to diagnose the difficulty of interpreting backtrackers and counterlegals in a single semantic framework. I have argued that a

[24] I take Glymour's *The Mind's Arrows* to be an elaboration and defense of an idea along these lines (Glymour 2001).

promising strategy for evaluating both backtrackers and counterlegals involves looking to our favored pictures of parts of the world and making minimal changes to the causal structures they depict. My strategy of argumentation has been to propose such a semantics.

**Acknowledgements** Thanks to Aldo Antonelli, Adam Edwards, Bas van Fraassen, Bernard Molyneux, and Ted Shear for their comments on earlier versions of this paper. Thanks to Alex Kocurek for some especially detailed and helpful comments on an earlier version. And most of all, thanks to my adviser, Adam Sennet, for all his help.

## References

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Clarendon Press.
Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, *160*, 139–166.
Fine, K. (1975). Review of Lewis' counterfactuals. *Mind*, *84*(335), 451–458.
Fisher, T. (2016). Causal counterfactuals are not interventionist counterfactuals. *Synthese*. doi:10.1007/s11229-016-1183-0.
Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *3*(1), 151–182.
Glymour, C. (2001). *The Mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT 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.
Lewis, D. (1973). *Counterfactuals*. Blackwell Publishing. First published in 1973.
Lewis, D. (1979a). Counterfactual dependence and time's arrow. *Noûs*, *13*, 455–476.
Lewis, D. (1979b). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*(1), 339–359.
Loewer, B. M. (1976). Counterfactuals with disjunctive antecedents. *Journal of Philosophy*, *73*, 531–537.
Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, *121*(1–2), 93–149.
Pearl, J. (2000). *Causality: Models, reasoning, and inferences* (1st ed.). Cambridge: Cambridge University Press.
Pollock, J. (1976). *Subjunctive reasoning*. Dordrecht: Reidel.
Pollock, J. (1981). A refined theory of counterfactuals. *Journal of Philosophical Logic*, *10*, 239–266.
Rips, L. J. (2009). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.
Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.
Tichý, P. (1976). A counterexample to the Stalnaker–Lewis analysis of counterfactuals. *Philosophical Studies*, *29*(4), 271–273.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.