Studies in Logic, Vol. 10, No. 1 (2017): 79–94
PII: 1674-3202(2017)-01-0079-16

Hiddleston's Causal Modeling Semantics and the Distinction between Forward-Tracking and Backtracking Counterfactuals*

Kok Yong Lee
Department of Philosophy, National Chung Cheng University
kokyonglee.mu@gmail.com

Abstract. Some cases show that counterfactual conditionals ('counterfactuals' for short) are *inherently ambiguous*, equivocating between forward-tracking and backtracking counterfactuals. Elsewhere, I have proposed a causal modeling semantics, which takes this phenomenon to be generated by two kinds of causal manipulations. (Lee 2015; Lee 2016) In an important paper (Hiddleston 2005), Eric Hiddleston offers a different causal modeling semantics, which he claims to be able to explain away the inherent ambiguity of counterfactuals. In this paper, I discuss these two semantic treatments and argue that my (bifurcated) semantics is theoretically more promising than Hiddleston's (unified) semantics.

1 Introduction

Jim is standing at a high cliff. What would have happened if Jim were to jump off the cliff? Naturally, there are two ways to counterfactualize the situation, and they give rise to two individually intuitive yet jointly incompatible verdicts. On the one hand, we may reason that Jim would have gotten killed by jumping off the cliff, since he would not be able to survive crashing on the ground after falling from such a high cliff. On the other hand, we may reason that Jim would not have gotten killed by jumping off the cliff, since Jim is a rational person, who will not jump off a high cliff unless there is, say, a safety net installed at the bottom. But if a safety net were installed at the bottom, Jim certainly would not have gotten killed by jumping off the cliff (he might even come out unhurt!). This shows that a counterfactual conditional (or 'counterfactual' for short) is *inherently ambiguous* in the sense that the same counterfactual, say, "If Jim were to jump off the cliff, he would have gotten killed" is true under one mode of counterfactualization but false under the other (also see [4]). Traditionally, these two modes of counterfactualization are regarded as resulting in

Received 2016-12-20

*A previous version of this paper had been presented in Workshop on Philosophical Logic: Conditionals and Related Questions at National Taiwan University. I want to thank all participants for their comments. I also want to thank Duen-Min Deng for helping me to improve the formulation of the causal modeling semantics presented here.

two kinds of counterfactuals, namely, *forward-tracking* and *backtracking* counterfactuals respectively. More precisely, the counterfactual "If Jim were to jump off the cliff, he would have gotten killed" is said to be true interpreted as a forward-tracking counterfactual, while false interpreted as a backtracking counterfactual.

The inherent ambiguity of counterfactuals, I have argued, is rooted in two distinct kinds of causal manipulation, which are responsible for the different ways of counterfactualizing exhibited in the example above. ([5]) It is for this reason that I have also suggested that the distinction between forward-tracking and backtracking counterfactuals is better characterized by the causal modeling semantics of counterfactuals.

In an important paper, Erick Hiddleston ([3]) has proposed a different causal modeling semantics of counterfactuals, which is claimed to be able to account for the inherent ambiguity of counterfactuals. Hiddleston's semantics is starkly different from the one I proposed before in that while my semantics appeals to distinct treatments of two types of counterfactuals, Hiddleston's semantics offers a unified treatment. In this paper, I want to compare and contrast these two semantic treatments. I argue that Hiddleston's unified treatment, unlike my bifurcated treatment, fails to capture the inherent ambiguity of counterfactuals.

In what follows, I will first introduce the causal modeling semantics I propose in earlier papers. I will then examine Hiddleston's semantics, and comparing his semantics with mine. I then point out the difficulties faced by Hiddleston's semantics.

2 Causal Modeling Semantics

Perhaps the best way to introduce the causal modeling semantics of counterfactuals is to look at a concrete example. Let us then construct a causal model J for the case mentioned at the beginning (I will call this case '*Jump*').

A causal model is a mathematical entity aiming at representing the causal relations of the events in a scenario. More formally, a causal model M is a quadruple $\langle U, V, S, A\rangle$. The first two elements, U and V, are sets of variables that are variables for events constituting the scenario that the causal model is supposed to represent. U is a finite set of variables $\{U_1, \ldots, U_n\}$ called the *exogenous variables*, which are supposed to be causally independent of all other factors in the model. V is a finite set of variables $\{V_1, \ldots, V_m\}$ called the *endogenous variables*, which are supposed to be causally dependent upon other factors in the model. The causal model J of *Jump* naturally contains the following endogenous variables:

JUMP represents whether or not Jim jumps off the cliff.
KILL represents whether or not Jim gets killed.

J also naturally contains the following exogenous variables:

RATIONAL represents whether or not Jim is a rational person.

NET represents whether or not a safety net is installed at the bottom of the cliff.

In general, each $V_i \in V$ and $U_i \in U$ admit a range of values, but it should be obvious that J only contains *binary variables* that take on two possible values, i.e., "Yes" or "No".

It is customary to use '$V_i = v_i$' to stand for the proposition *The variable V_i takes on the value of v_i*. For binary variables such as JUMP, KILL, RATIONAL, NET, we may use '1' and '0' to stand for Yes and No respectively (for simplicity's sake, this paper will only deal with binary variables). For instance, "JUMP $= 1$" means that Jim jumps off the cliff, while "NET $= 0$" means that no safety net is installed at the bottom of the cliff.

The third element of a causal model, S, is a set of *structural equations* that specify the causal-dependence relationships among variables. The causal dependence in play may be deterministic and indeterministic, although I will focus solely on deterministic causal relations here. For each $V_i \in V$, S contains exactly one structural equation of the following form:

$$V_i \Leftarrow f_i(\text{PA}_i).$$

The meaning of '$\Leftarrow$' is twofold. On the one hand, "$X \Leftarrow Y$" means that X is causally dependent on Y, i.e., whether X obtains or not is causally dependent on whether Y obtains or not. On the other hand, "$X \Leftarrow Y$" indicates that X will take on the value of Y. Let 'PA_i' stand for a subset of $U \cup V$ which is the set of V_i's *parents*. Parenthood is essentially a causal relation: the parents of an event are its direct causes, and its children are its direct effects. The parents of a variable occur in the right-hand side of its structural equation. For simplicity's sake, we will also treat variables on the right-hand side of the equation as propositions such that "Y" means $Y = 1$, and "$\sim Y$" means $Y = 0$.

J's S naturally contains the following structural equations:

$$\text{JUMP} \Leftarrow (\sim\text{RATIONAL} \vee \text{NET})$$
$$\text{KILL} \Leftarrow (\text{JUMP} \wedge \sim\text{NET})$$

In words, "JUMP $\Leftarrow$ ($\sim$RATIONAL $\vee$ NET)" means that whether or not Jim jumps off the cliff depends causally on both whether or not Jim is a rational person and whether or not a safety net is installed at the bottom such that Jim will jump off the cliff *if and only if* either he is irrational or a safety net is installed at the bottom. "KILL $\Leftarrow$ (JUMP $\wedge \sim$NET)" means that whether or not Jim gets killed is causally dependent on both whether or not Jim jumps off the cliff and whether or not a safety net is installed such that Jim will get killed if and only if he jumps off the cliff and there is no safety net installed at the bottom.

There is no structural equation for exogenous variables such as RATIONAL and NET. For exogenous variables are assumed to be causally independent of all other

factors in the model. Their values are "given" in the model rather than determined by the structural equations.

The fourth element of a causal model, A, is a function that assigns values to all variables in the model. J's A, arguably, is as follows:

$$\begin{aligned} A(\text{RATIONAL}) &= 1 \\ A(\text{NET}) &= 0 \\ A(\text{JUMP}) &= 0 \\ A(\text{KILL}) &= 0. \end{aligned}$$

In words, in *Jump*, Jim is a rational person, there is no safety net installed down the cliff, Jim does not jump off the cliff, and he does not get killed.

With the notion of causal model at hand, we are in a position to introduce the causal modeling semantics. At its core, the semantics takes the truth condition of counterfactuals as:

(CM) "$A > C$" is true in a causal model M iff "C" is true in certain submodels M'.

'$>$' stands for the counterfactual-conditional connective. Informally, a submodel M' is a causal model generated by causally manipulating M in a certain way. The general idea behind CM is this. Since a causal model M represents a scenario s, a counterfactual scenario s', generated by causally manipulating the scenario s, is represented by a submodel M' of M, which is generated in turn by causally manipulating M in a parallel way.

I have argued, in previous works, that there are two types of submodels, which give rise to the distinction between forward-tracking and backtracking counterfactuals. ([4, 5]) The idea is that there are two distinct kinds of causal manipulation. Roughly, one may manipulate a causal model either by changing the value of a variable through breaking some structural equations or by changing the value of a variable through tracing the required modifications back to some exogenous variables. Let us call them *intervention* and *extrapolation* respectively.

Intervention has been featured in all prominent causal modeling semantics of counterfactuals ([2, 6, 1]). Let $M = \langle U, V, S, A\rangle$ be a causal model, B be a sentence of the form '$C_1 = c_1 \wedge \ldots \wedge C_n = c_n$', *VB* be the set of variables that are in B. An *intervention in M with respect to B* generates a submodel $M_{(B)} = \langle U_{(B)}, V_{(B)}, S_{(B)}, A_{(B)}\rangle$ of M such that:

(i) $U_{(B)} = U \cup VB$.
(ii) $V_{(B)} = V \setminus VB$.
(iii) $S_{(B)} = S$ except that for each $C_i \in VB \cup U_{(B)}$, $S_{(B)}$removes the structural equation $C_i \Leftarrow f_i(\text{PA}_i)$ from S.
(iv) $A_{(B)} = A$ except that for each $C_i \in VB \cap U_{(B)}$, $A_{(B)}$ sets the value of C_i to c_i.

To intervene in a causal model M with respect to B is to remove the original structural equations (if any) for $C_i \in VB$ and directly set the value to be c_i. If C_i is exogenous, intervention simply sets the value of C_i to be c_i. The values of the rest of the variables are calculated based on the value of C_i and $S_{(B)}$.

Extrapolation, by contrast, has received little attention from philosophers. Suppose $M = \langle U, V, S, A\rangle$ is a causal model, B a sentence of the form '$C_1 = c_1 \wedge \ldots \wedge C_m = c_m$', and *VB* the set of variables that are in B. Define VB^c to be the *closure* of the parents of the variables in *VB*, i.e., the set of the 'ancestors' of *VB*. That is to say, define VB^c to be the smallest set that satisfies the following conditions:

(i) For any $X \in VB$, $X \in VB^c$.

(ii) For any X and Y in $U \cup V$, if $X \in VB^c$ and Y is a parent of X (i.e., Y occurs in the structural equation for X), then $Y \in VB^c$.

Now, let $M = \langle U, V, S, A\rangle$ be a causal model and M^* a *submodel of M generated by extrapolating M with respect to B*, if M^* satisfies the following conditions:

(i) $U^* = U$.

(ii) $V^* = V$.

(iii) $S^* = S$.

(iv) $A^* = A(X)$ for each $X \in U \setminus VB^c$ and $A^*(C_i) = c_i$ for each C_i in *VB*.

Like intervention, to extrapolate a causal model M with respect to B also sets each C_i in *VB* to take on the value c_i. But unlike intervention, extrapolation preserves the structural equations of the original model. More importantly, while intervention always gives us a unique submodel, extrapolation may generate multiple submodels. When more than one submodel is generated, the context will determine which submodel or submodels are relevant in determining the truth values of counterfactuals. Let us use $\mathbf{M}^{(B)}$ to denote the contextually determined submodel or submodels M^*, which are generated by extrapolating M with respect to B, and which play a crucial role determines the truth value of the counterfactuals in play.

With intervention and extrapolation in hand, we may *disambiguate* CM into:

(CM_{IN}) "$A > C$" is true_{IN} in M iff "C" is true in $M_{(B)}$.

(CM_{EX}) "$A > C$" is true_{EX} in M iff "C" is true in all $M^* \in \mathbf{M}^{(B)}$.

CM_{IN} and CM_{EX} give the correct verdicts with respect to *Jump*. Intervening in J with respect to $(\text{JUMP} = 1)$ gives rise to the submodel $J_{(\text{JUMP}=1)}$ such that $J_{(\text{JUMP}=1)}$'s $U_{(\text{JUMP}=1)}$ and $V_{(\text{JUMP}=1)}$ are identical to J's. $J_{(\text{JUMP}=1)}$'s $S_{(\text{JUMP}=1)}$, by contrast, consists of the following:

$$\text{KILL} \Leftarrow (\text{JUMP} \wedge \sim\text{NET})$$

As a result, $J_{(\text{JUMP}=1)}$'s $A_{(\text{JUMP}=1)}$ is that:

$$\begin{aligned} A_{(\text{JUMP}=1)}(\text{RATIONAL}) &= 1 \\ A_{(\text{JUMP}=1)}(\text{NET}) &= 0 \\ A_{(\text{JUMP}=1)}(\text{JUMP}) &= 1 \\ A_{(\text{JUMP}=1)}(\text{KILL}) &= 1. \end{aligned}$$

On CM_{IN}, since "KILL $= 1$" is true in $J_{(\text{JUMP}=1)}$, "JUMP $= 1 >$ KILL $= 1$" is true_{IN} in J, as desired.

By contrast, suppose that we extrapolate J with respect to (JUMP $= 1$). In the present context, extrapolation arguably generates a unique submodel $J^{(\text{JUMP}=1)} \in \mathbf{J}^{(\text{JUMP}=1)}$, whose value assignment $A^{(\text{JUMP}=1)}$ is as follows:

$$\begin{aligned} A^{(\text{JUMP}=1)}(\text{RATIONAL}) &= 1 \\ A^{(\text{JUMP}=1)}(\text{NET}) &= 1 \\ A^{(\text{JUMP}=1)}(\text{JUMP}) &= 1 \\ A^{(\text{JUMP}=1)}(\text{KILL}) &= 0. \end{aligned}$$

On CM_{EX}, since "KILL $= 1$" is false in $J^{(\text{JUMP}=1)}$, "JUMP $= 1 >$ KILL $= 1$" is false_{EX} in J, as desired.

Not only does the distinction between intervention and extrapolation give the correct verdicts, it also sheds an important light on the two modes of counterfactualization that give rise to forward-tracking and backtracking counterfactuals. Let 'Jump' and 'Kill' stand for the propositions *Jim jumps off the cliff* and *Jim has gotten killed* respectively. When counterfactualizing that "Jump > Kill" is true in *Jump*, we focus solely on the causal effect of the event of Jim jumping off the cliff (i.e., Jump) itself. The causal relations between Jump and its causes are ignored. In particular, we make no attempt to actualize or rationalize how Jump could have happened in *Jump* in the first place. For instance, we ignore the facts that Jim is a rational person and that no safety net is installed, which in the actual situation have prevented Jim from jumping off the cliff. In a sense, we simply stipulate that Jim jumps off the cliff without having in our mind a specific story as to how Jump could have happened in the first place. This mode of counterfactualization is nicely captured by intervention, for intervening in a causal model M with respect to $(C_i = c_i)$ generates a submodel $M_{(C_i=c_i)}$ that contains information necessary for understanding the causal effect of $(C_i = c_i)$. ([2]) $M_{(C_i=c_i)}$ surgically removes the causal influence C_i's parents have on C_i, while stipulating C_i to take on the value c_i. This allows us to see clearly the causal effect that $(C_i = c_i)$ has on C_i's children.

On the other hand, when counterfactualizing that "Jump > Kill" is false in *Jump*, our focus is on the causal relations among Jump and its causes in order to determine

under what condition Jump could have actually happened. For instance, we reason that Jim would not get killed if he was to jump off the cliff, since Jim was a rational person, and a rational person would not jump off the cliff without the installation of a safety net at the bottom, but if a safety net was installed, jumping off the cliff would then not get him killed. This mode of counterfactualization is captured nicely by extrapolation, as extrapolating a causal model M with respect to $(C_i = c_i)$ generates a set of submodels M^* that contains all information necessary for knowing under what condition $(C_i = c_i)$ could have actually happened in M. M^* assigns the values of its variables in a way that preserves all the causal relations among its variables in M, which gives us a story of what else needs to change in order for C_i to take on the value c_i in M.

3 Hiddleston's Causal Modeling Semantics

In [3], Eric Hiddleston proposes a different causal modeling semantics of counterfactuals which is in stark contrast to the one introduced above. Specifically, Hiddleston's semantics offers a unified account of forward-tracking and backtracking counterfactuals. In what follows, I first will introduce Hiddleston's semantics, pointing out the similarities and differences between Hiddleston's semantics and the one mentioned above (or the orthodox causal modeling semantics in general). I then argue that Hiddleston's semantics fails to account for the distinction between forward-tracking and backtracking semantics. Rather, closely examining what goes wrong in Hiddleston's semantics further vindicates the assumption that forward-tracking counterfactuals and backtracking counterfactuals are of two different kinds.

A distinctive feature of Hiddleston's semantics is that it allows indeterministic laws. More precisely, Hiddleston takes structural equations to be specified as follows:

(H) $\Box((A_1 = a_1 \wedge \ldots \wedge A_n = a_n) \supset \Pr(C = c) = x)$.

'$\supset$' and 'Pr' stand for material implication and the probability function respectively. Hiddleston restricts A_i to what he calls the *positive parents* of C in M. Positive parenthood characterizes the variables which have a *direct positive influence* on $C = c$. The latter is defined as follows:

For each A_i, A_i has a direct positive influence on $C = c$ iff $\Pr(C = c|A_i = a_i \wedge Z_i = z_i) > \Pr(C = c|A_i \neq a_i \wedge Z_i = z_i)$ (where Z_i stands for C's other parents).

We now define a kind of submodel M' of M which Hiddleston calls "Φ-minimal model'. To get to it, we need to introduce some terminologies.

As noted, a submodel M' of M is a causal model resulted from causally manipulating M in some specific manners (M' and M would thus have the same set of variables V and U). A Φ-model is a causal model in which "Φ" is true.

Let '$\mathrm{PPA_{C,M}}$' stand for the set of C's positive parents in M such that $\mathrm{PPA_{C,M}} =$

$\{A_i : A_i = a_i$ has a direct positive influence on $C = c$ in $M\}$. When no confusion arises, we may drop the subscript of M.

A *causal break* is a variable, whose value in a submodel M' is different from its value in M while all its positive parents have the same values in M' as in M. More precisely, a causal break in a submodel M' relative to M is a variable A such that the value of A in M' is different from the value of A in M, and for each $X_i \in \text{PPA}_\text{A}$, the value of X_i remains constant across M' and M. Let 'Break(M', M)' be the set of variables A_i such that A_i is a causal break in M' relative to M. When no confusion arises, we may simply write 'Break'.

A *causal intact* is a variable, whose value in a submodel M' is the same as the one in M and all its positive parents have the same values in M' as in M. More precisely, a causal intact in a submodel M' relative to M is a variable A such that the value of As remains constant across M' and M, and for each $X_i \in \text{PAA}_\text{A}$, the value of X_i remains constant across M' and M. Let 'Intact(M', M)' be the set of variables A_i such that A_i is a causal intact in M' relative to M. When no confusion arises, we may simply write 'Intact'.

Now, we are in a position to define a Φ-minimal model M' which is crucial to Hiddleston's account. Let $M = \langle V, U, S, A\rangle$ be a causal model, A submodel M' of M and Break(M', M) are Φ-minimal relative to M iff

(i) M' is a Φ-model,
(ii) For each variable X which is not a descendant of Φ, Intact$(M', M) \cap \{X\}$ is maximal among Φ-models, and,
(iii) Break(M', M) is minimal among Φ-models.

Hiddleston's causal modeling characterization of the truth condition of counterfactuals is as follows:

(CM_H) "$A > C$" is true in a model M and a context C iff "C" is true in every A-minimal model M' for which Break(M', M) is relevant in C.

Notice that CM_H relates the truth-values of counterfactuals to contexts. The reason is that there may be multiple (yet incompatible) A-minimal models M', and only the relevant A-minimal model is pertaining to the characterization of the truth condition of "$A > C$", while whether a A-minimal model is relevant is determined by context. When no confusion arises, we will drop the specification of the context.

Before we go on to discuss Hiddleston's treatment of forward-tracking and backtracking counterfactuals, let us pause and make some comments. First, a distinctive feature of Hiddleston's semantics is that it allows structural equations to be specified by a probabilistic function, i.e., (H). Hiddleston's idea is that (H) embodies a *quasi-deterministic* view on causal dependence: an event A is causally dependent on an event B even if B only renders A more probable rather than certain. Hiddleston justifies (H) by pointing out that "many processes such as coin flips and die

rolls behave as if they were indeterministic, and so we treat them as such". "This quasi-determinism," Hiddleston contends, "may be due to either determinism or indeterminism at the fundamental level, and commonsense is not committed to either way". ([3], p. 639)

Nevertheless, Hiddleston's quasi-deterministic structural equations can account for the orthodox deterministic structural equations that take the form $V_i \Leftarrow f_i(\mathrm{PA}_i)$. For notice that the following is a special case of (H):

$$(\mathrm{H_D})\ \Box((A_1 = a_1 \wedge \ldots \wedge A_n = a_n) \supset \Pr(C = c) = 1).$$

Now, we may further reformulate ($\mathrm{H_D}$) into:

$$(\mathrm{H_D})'\ C \Leftarrow f_i(A_1 = a_1 \wedge \ldots \wedge A_n = a_n).$$

F_i is a certain (causal) function that maps $(A_1 = a_1 \wedge \ldots \wedge A_n = a_n)$ to c. That ($\mathrm{H_D}$) and ($\mathrm{H_D}$)$'$ are basically the same is warranted by strict implication.

It is an interesting question whether we should adopt quasi-deterministic structural equations as Hiddleston does or deterministic structural equations as the orthodox causal modeling semanticists do. While I agree with Hiddleston that commonsense is not committed to either determinism or indeterminism, it is not obvious to me that quasi-deterministic causal dependence is ubiquitous in our understanding of daily situations. The reason is that our understanding of a situation often consists in grasping the *circumstantial necessity*, i.e., what is inevitable in the circumstance, among events. For instance, in *Jump*, it is true that, strictly, Jim may not even get hurt jumping off a high cliff. So Jim getting killed is only quasi-deterministically depends on Jim jumping off the cliff. Yet it is common that we idealize the situation so that Jim getting killed is circumstantially inevitable given that he jumps off the cliff. Such idealization is understandable and even mandatory, for otherwise many situations would not be graspable. Hiddleston is surely right that processes such as coin flips and die rolls are characteristically quasi-deterministic. The orthodox causal modeling semanticists, however, can always handle such processes by regarding them as exogenous variables.

Undoubtedly, a lot more can and should be said regarding this issue. Pursuing the issue any further, however, is beyond the scope of this paper. Fortunately, the point I want to make will not be affected by our choice of the general form of structural equations. For a deterministic structural equation can be regarded as a special case of the quasi-deterministic form of structural equations, and my argument can be manifested by using only the deterministic structural equations.

Second, suppose that a causal model M contains only structural equations of the form ($\mathrm{H_D}$)$'$. It follows that a Break related to M can only be an exogenous variable. For it is impossible for an endogenous variable to take a different value while its parent's value remains intact, given that the structural equation in play is of determinism. Moreover, with respect to such a model, a set of A-minimal models is identical to a

certain set of submodels $\mathbf{M}^{(A)}$. For it seems obvious that an appropriate specification will allow an extrapolation of M with respect to A to satisfy the three conditions of A-minimal model mentioned above. In other words, $\mathrm{CM_H}$ can be characterized by $\mathrm{CM_{EX}}$, when only deterministic structural equations (i.e., $(\mathrm{H_D})'$) are involved.

4 Hiddleston on Forward-Tracking and Backtracking Counterfactuals

Hiddleston does not take counterfactuals to be inherently ambiguous in the sense defined above. Rather he takes the distinction between forward-tracking and backtracking counterfactuals to be manifested by a certain context-dependent feature of counterfactuals. As noted, $\mathrm{CM_H}$ takes the truth condition of "$A > C$" to be relative to a certain Break determined by a certain context. Such a context-dependence of the truth condition of counterfactuals, on Hiddleston's view, results in the distinction between forward-tracking and backtracking counterfactuals. Let me elaborate.

Arguably, Hiddleston will accept J as "a natural model" for *Jump*. ([3], p. 645) Firstly, that J contains RATIONAL, NET, JUMP, KILL seems both natural and intuitive. Secondly, it should be uncontroversial that J's S consist of:

$$\mathrm{JUMP} \Leftarrow (\sim\mathrm{RATIONAL} \vee \mathrm{NET})$$
$$\mathrm{KILL} \Leftarrow (\mathrm{JUMP} \wedge \sim\mathrm{NET}).$$

For one thing, we have seen that (H) can be construed as $(\mathrm{H_D})'$, when only deterministic structural equations are involved. For another, Hiddleston also notes that in such a case, "Jim jumps only if either NET $= 1$ or RATIONAL $= 0$" ([3], p. 645; I have modified Hiddleston's remarks to be in line with the present terminology). Thirdly, J's value assignment A is also as innocuous as it can be, reflecting the fact that Jim is a rational person, who does not jump off a high cliff without a safety net installed at the bottom. Hiddleston has accepted A. ([3], p. 645)

I have claimed that "If Jim were to jump off the cliff, he would have gotten killed" (or "JUMP $= 1 >$ KILL $= 1$") is true when construed as a forward-tracking counterfactual but false when construed as a backtracking counterfactual. As I see it, Hiddleston also agrees with this claim. However, Hiddleston does not think that the difference between these two kinds of counterfactuals consists in two different kinds of causal manipulations. The difference, rather, is considered as the product of the context-sensitivity of the relevant Break. ([3], pp. 645–646) On Hiddleston's view, "JUMP $= 1 >$ KILL $= 1$" is true when $\{$RATIONAL$\}$ is taken as the relevant Break, whereas it will become false when $\{$NET$\}$ is taken as the relevant Break instead. More precisely, when $\{$RATIONAL$\}$ is taken to be the relevant Break, the only JUMP $= 1$-

minimal model J' is such that J''s A' is as follows:

$$\begin{aligned} A'(\text{RATIONAL}) &= 0 \\ A'(\text{NET}) &= 0 \\ A'(\text{JUMP}) &= 1 \\ A'(\text{KILL}) &= 1. \end{aligned}$$

By contrast, when $\{\text{NET}\}$ is regarded as the relevant Break, the only JUMP $= 1$-minimal model J'''s A'' is as follows:

$$\begin{aligned} A''(\text{RATIONAL}) &= 1 \\ A''(\text{NET}) &= 1 \\ A''(\text{JUMP}) &= 1 \\ A''(\text{KILL}) &= 0. \end{aligned}$$

CM_H gives verdicts that are in accordance with our initial intuitions. On the one hand, since "KILL $= 1$" is true in J' which is the only JUMP $= 1$-minimal model for which Break(J, J') (i.e., $\{\text{RATIONAL}\}$) is relevant, "JUMP $= 1 >$ KILL $= 1$" is true in J, as desired. On the other hand, since "KILL $= 1$" is false in J'', which is the only JUMP $= 1$-minimal model for which Break(J, J'') (i.e., $\{\text{NET}\}$) is relevant, "JUMP $= 1 >$ KILL $= 1$" is false in J, as desired. The variation of the truth-value of "JUMP $= 1 >$ KILL $= 1$" is regarded as the product of the context-sensitivity of Break.

Which one should we choose, a unified treatment such as CM_H, or a bifurcated one such as CM_IN and CM_EX? The key to this question is intervention. It is not hard to recognize that Hiddleston's semantics in general leaves no room for intervention. For all A-minimal models M' of M preserve the set of structural equations of M, and without the violation of certain structural equations, intervention is impossible. This suggests a natural way to test Hiddleston's treatment of the ambiguity of counterfactuals. That is, in order for CM_H to hold, or at least be theoretically no less promising than CM_IN and CM_EX, it must be that the distinction between forward-tracking and backtracking counterfactuals can always be explained or predicted by the context-sensitivity of Break. But this last point is problematic. One way to see this is to note that, related to causal models containing only structural equations of the form $(\text{H}_\text{D})'$ and binary variables, the context-sensitivity-of-Break maneuver is feasible only when there are more than one exogenous variables, otherwise there will only be at most one A-minimal model M' of M. In other words, such a causal model will only have exactly one Break, i.e., the only exogenous variable, and, as a result, the truth-values of counterfactuals with respect to such a model could not be context-sensitive. The problem is that the distinction between forward-tracking and backtracking counterfactuals

persists even in causal models with exactly one exogenous variable. For instance, suppose that we modified *Jump* such that either a powerful demon will install a safety net at the bottom or she will cause Jim to become a rational person (call this case '*Jump**'). Naturally, a causal model J^* for *Jump** contains exactly one exogenous variable:

> DEMON represents whether the demon installs a safety net at the bottom of the cliff or she causes Jim to become a rational person.

By contrast, J^*'s endogenous variables include NET, RATIONAL, JUMP, KILL. The detail of J^* needs not bother us here. What is important is while J^* does not allow for more than one JUMP $=$ 1-minimal model. But it seems that the distinction between construing "JUMP $=1>$ KILL $=1$" as a forward-tracking counterfactual and construing it as a backtracking counterfactual is still perfectly sensible in *Jump**. Specifically, it seems that "JUMP $=1>$ KILL $=1$" still appears to be true (false) when construed as a forward-tracking (backtracking) counterfactual in *Jump**.

While it is not hard to see that $\mathrm{CM_{IN}}$ and $\mathrm{CM_{EX}}$ can give the desired verdicts for the truth-values of "JUMP $=1>$ KILL $=1$" in *Jump**, the same cannot be said of $\mathrm{CM_H}$. For models that contain exactly one exogenous variable like J^*, $\mathrm{CM_H}$ will unduly predict either that the distinction between forward-tracking and backtracking counterfactuals does not arise, or that these two kinds of counterfactuals collapse. Neither option seems plausible. This shows not only that $\mathrm{CM_H}$ is not in a position to account for the inherent ambiguity of counterfactuals, but also that a bifurcated treatment along the line of $\mathrm{CM_{IN}}$ and $\mathrm{CM_{EX}}$ is on the right track.

The problem is further manifested by cases where intervention and extrapolation come apart. In the extreme cases, there can be intervention even if no extrapolation is possible. To illustrate, consider the following case:

> *Nuclear*. A nuclear missile will be launched if two separate passcodes are keyed into the launching machine. If the missile launches, a major city will be destroyed. The captain is the only person who knows both passcodes. If he decides to launch the missile, then he will have to give each of his two assistants, John and Jason, a separate passcode, and they will then key it into the launching machine. The captain has no intention to destroy any city. To make sure that the missile will not be launched, the captain hypnotizes himself such that he will be psychologically impossible to give both John and Jason a passcode. However, the laws require that the captain have to tell at least one of his assistants one of the two passcodes. The captain tells John the passcode.

Let us construct a causal model N for *Nuclear*. N's U naturally contains one exogenous variable:

> CAPTAIN represents whether the captain gives a passcode to John or to Jason.

Also, we stipulate that CAPTAIN takes on the value 1 when the captain decides to give a passcode to John, otherwise the value 0.

N's V, by contrast, consists of four endogenous variables:

> JOHN represents whether or not John keys a passcode into the launching machine.
> JASON represents whether or not Jason keys a passcode into the launching machine.
> LAUNCH represents whether or not the nuclear missile is launched.
> DESTROY represents whether or not a major city is destroyed.

The following are the structural equations in N's S:

$$\begin{aligned} \text{JOHN} &\Leftarrow \text{CAPTAIN} \\ \text{JASON} &\Leftarrow \sim\text{CAPTAIN} \\ \text{LAUNCH} &\Leftarrow (\text{JOHN} \wedge \text{JASON}) \\ \text{DESTROY} &\Leftarrow \text{LAUNCH}. \end{aligned}$$

In words, whether John (Jason) keys in the passcode depends causally on whether or not the captain tells him the passcode such that John (Jason) will key in the passcode if and only if the captain tells him the passcode. Moreover, whether or not the nuclear missile will launch depends causally on whether or not John and Jason key in the passcode such that the missile will launch if and only if both John and Jason key in the passcode. Finally, whether or not a major city will be destroyed depends causally on whether or not the nuclear missile launches such that the city will be destroyed if and only if the missile launches.

Naturally, N's A is as follows:

$$\begin{aligned} A(\text{CAPTAIN}) &= 1 \\ A(\text{JOHN}) &= 1 \\ A(\text{JASON}) &= 0 \\ A(\text{LAUNCH}) &= 0 \\ A(\text{DESTROY}) &= 0. \end{aligned}$$

In words, the captain tells John the passcode, John keys in the passcode, Jason does not key in the passcode, the nuclear missile does not launch, and a major city is not destroyed.

N shows that intervention and extrapolation cannot be the same. More precisely, while there is a solution when intervening in N with respect to $(\text{LAUNCH} = 1)$, there is no solution when extrapolating N with respect to $(\text{LAUNCH} = 1)$. That is, intervening in N with respect to $(\text{LAUNCH} = 1)$ generates a submodel $N_{(\text{LAUNCH}=1)}$

whose set of structural equations $S_{(\text{LAUNCH}=1)}$ consists of:

$$\begin{aligned} \text{JOHN} &\Leftarrow \text{CAPTAIN} \\ \text{JASON} &\Leftarrow \sim\text{CAPTAIN} \\ \text{DESTROY} &\Leftarrow \text{LAUNCH}. \end{aligned}$$

Moreover, $A_{(\text{LAUNCH}=1)}$ is as follows:

$$\begin{aligned} A(\text{CAPTAIN}) &= 1 \\ A(\text{JOHN}) &= 1 \\ A(\text{JASON}) &= 0 \\ A(\text{LAUNCH}) &= 1 \\ A(\text{DESTROY}) &= 1. \end{aligned}$$

By contrast, extrapolating H with respect to $(\text{LAUNCH} = 1)$ generates no consistent submodel N^* at all. Suppose that we extrapolate N with respect to $(\text{LAUNCH} = 1)$. By $\text{LAUNCH} \Leftarrow (\text{JOHN} \wedge \text{JASON})$, it follows that JASON should take on the value 1. But then CAPTAIN will have to take on the value 0 (by $\text{JASON} \Leftarrow \sim\text{CAPTAIN}$). But if CAPTAIN is to take on the value 0, JOHN also will take on the value 0 (by $\text{JOHN} \Leftarrow \text{CAPTAIN}$). But if JOHN is to take on the value 0, LAUNCH will have to take on the value 0, too (by $\text{LAUNCH} \Leftarrow (\text{JOHN} \wedge \text{JASON})$). Contradiction. In other words, extrapolation N with respect to $(\text{LAUNCH} = 1)$ will have no solution.

Since CM_H can be characterized by CM_EX related to causal models like N, it is not surprising that the former, too, is not able to handle the same problem. Notice that in N, the only relevant break is $\{\text{CAPTAIN}\}$. But if so, then there exists no $\text{LAUNCH} = 1$-minimal model N' for which $\{\text{CAPTAIN}\}$ is relevant. LAUNCH to take on the value 1 is impossible in the sense that it requires breaking structural equations. This indicates the root of this problem: since both CM_H and CM_EX do not allow violations of structural equations, some value assignments may thus turn out impossible.

This is problematic if CM_H is supposed to account for forward-tracking counterfactuals. In particular, the following (forward-tracking counterfactual) seems intuitively true in *Nuclear*:

(1) If the nuclear missile had been launched, a major city would have been destroyed.

(1) causes no problem for my semantics, for "DESTROY=1" is true_IN in $N_{(\text{LAUNCH}=1)}$ as desired. But the same could not be said of CM_H, as we can see that it is impossible for LAUNCH to take on value 1 for doing so requires violations of structural equations. Hiddleston, in a footnote, suggests taking such counterfactuals to be vacuously true. ([3], p. 655, footnote 7) So perhaps we can regard (1) as vacuously true. But this

move is implausible, for the following (forward-tracking counterfactual) would also be counted as vacuously true:

(2) If the nuclear missile had been launched, a major city would still not have been destroyed.

Since (2) is intuitively false, Hiddleston's suggestion is implausible.

5 Conclusion

If what have been said is correct, Hiddleston's causal modeling semantics cannot cope with the inherent ambiguity of counterfactuals. While such ambiguity might sometimes be predicted by the context-sensitivity of the relevant Breaks in CM_H, it is mistaken to diagnose the root of this phenomenon as consisting in such context-sensitivity. Elaborating the failure of CM_H in fact shows clearly that a bifurcated semantics such as CM_{IN} and CM_{EX} is required in order to account for explaining the inherent ambiguity of counterfactuals. Hence, contra Hiddleston, intervention and extrapolation are the key to the distinction between forward-tracking and backtracking counterfactuals.

References

[1] R. Briggs, 2012, "Interventionist counterfactuals", *Philosophical Studies*, **160(1)**: 139–166.

[2] D. Galles and J. Pearl, 1998, "An axiomatic characterization of causal counterfactuals", *Foundations of Science*, **3(1)**: 151–182.

[3] E. Hiddleston, 2005, "A causal theory of counterfactuals", *Noûs*, **39(4)**: 632–657.

[4] K. Y. Lee, 2015, "Causal models and the ambiguity of counterfactuals", in W. van der Hoek, W. H. Holliday and W.-F. Wang (eds.), *Logic, Rationality, and Interaction: 5th International Workshop, LORI 2015*, pp. 220–229, New York: Springer.

[5] K. Y. Lee, 2016, "Motivating the causal modeling semantics of counterfactuals, or, why we should favor the causal modeling semantics over the possible-worlds semantics", in S. C.-M. Yang, D.-M. Deng and H. Lin (eds.), *Structural Analysis of Non-Classical Logics: The Proceeedings of the Second Taiwan Philosophical Logic Colloquium*, pp. 83–110, New York: Springer.

[6] J. Pearl, 2000, *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.

Hiddleston 因果模型语义学以及前进式与回溯式反事实条件句的区别

李国扬
国立中正大学 哲学系
kokyonglee.mu@gmail.com

摘　要

某些案例显示反事实条件句（counterfactual conditionals）是有“内在歧义的”（inherently ambiguous），即同一句反事实条件句既可以表达“前进式反事实条件句”（forward-tracking counterfactuals）也可以表达“回溯式反事实条件句”（back-tracking counterfactuals）。在之前的文章中（Lee 2015，Lee 2016），我提出一个因果模型反事实条件句语义学（causal modeling semantics of counterfactuals），主张反事实条件句的内在歧义性是由不同的因果操弄（causal manipulation）所产生的。在一篇很重要的论文中（Hiddleston 2005），Eric Hiddleston 提出一个截然不同的因果模型反事实条件句语义学，并宣称这个语义学可以解释反事实条件句的内在歧义性。本文将介绍上述两个因果模型反事实条件句语义学，并试图论证本人的语义学比 Hiddleston 的语义学能够更好地处理反事实条件句的内在歧义性。