Backtracking Counterfactuals Revisited

Backtracking counterfactuals, revisited

Justin Khoo

Forthcoming in *Mind*

Abstract

I discuss three observations about backtracking counterfactuals not predicted by existing theories, and then motivate a theory of counterfactuals that does predict them. On my theory, counterfactuals quantify over a suitably restricted set of historical possibilities from some contextually relevant past time. I motivate each feature of the theory relevant to predicting our three observations about backtracking counterfactuals. The paper concludes with replies to three potential objections.

Consider the following situation (cf. Jackson 1977): you see your friend Smith on the roof of a twenty story building, poised to jump. There is nothing underneath him besides the solid concrete of the sidewalk. You feel anxiety and fear—you do not want your friend to die! Trying to regain composure, you remind yourself that you know Smith and know that he is rational, has no wish to die, and knows that (since there is nothing underneath him) jumping in such circumstances will kill him. Reflecting on this, you think: he is not going to jump. Thankfully, just as you predict, Smith steps down off the ledge and descends the stairs, exiting the building safely. Relieved, you say, 'Thank goodness,

(1) If Smith had jumped, he would have died.'

It seems clear that the counterfactual you utter is true. Furthermore, that (1) is true seems to be why it is appropriate to feel relief when Smith does not jump; it also seems to be why your anxiety that he might jump is reasonable in light of the circumstances.[1]

Now, Beth is also on the scene, and hears you utter (1). Beth objects on the following grounds. 'Smith was rational, had no wish to die, could see below him, and knew that

[1] I have in mind something like the following: feeling relief that $\neg p$ is appropriate iff things would have been worse off had p occurred (or perhaps iff you believe things would have been worse off had p occurred). Some sort of counterfactual comparison seems to be a property of emotive factive verbs generally (compare *regret, resent*) and may be related to the fact that emotive factious seem to presuppose knowledge (not just truth) of their complements—see for instance Zuber 1977.

jumping without a net would kill him,' Beth says. 'Therefore, had Smith jumped, there would already have been a net below him to catch him safely.[2] Hence,

(2) If Smith had jumped, he would have lived.'

I submit that Beth has made a pretty good case against (1) and for (2)—though perhaps one that would not convince you that your earlier anxiety was misplaced. Maybe you say: 'I know all those things about Smith. But there was no net underneath him! Hence, (1).' Perhaps Beth remains unconvinced, and the discussion continues for several more rounds.

Backtracking counterfactuals have been discussed, often as an oddity to be set aside, in many places.[3] The aim of this paper is to give backtracking counterfactuals their due. Here is the plan for what follows. In §1, I draw three important observations about backtracking and non-backtracking counterfactuals, observations which a plausible theory of the meaning of counterfactuals ought to predict. In §2, I draw on the three observations to raise a challenge to the most well-known 'similarity theory' of counterfactuals (e.g., Stalnaker 1968, Lewis 1973a, 1979a). The challenge is to predict these observations without making ad hoc stipulations, and I know of no existing theory that does so. In §3, I turn to an alternative theory of counterfactuals on which they quantify over alternative histories 'branching' from the actual world at some past time.[4] According to my favored version of the 'historical modality' theory, there are two contextual parameters relative to which the truth of a counterfactual is evaluated: a time, and a set of salient propositions. In §4, I combine this theory with two independently motivated pragmatic principles of speaker interpretation; these principles predict default settings for these two contextual parameters and also allow that in certain conditions these defaults may be overridden. I show that the resulting theory predicts our three observations. Finally, in §5, I consider how my theory compares with David Lewis's similarity theory and respond to several objections. Although my main goal is to articulate a plausible theory of counterfactuals that predicts our four observations about backtracking counterfactuals, a related secondary goal is to illustrate the usefulness of adopting a historical modality theory of counterfactuals, thus

[2] Or, 'had Smith jumped, there would have to have been a net below him to catch him safely,' or 'Smith would have jumped only if there had been a net below him to catch him safely'.

[3] See for instance, Downing 1959, Jackson 1977, Lewis 1979a, Bennett 1984, 2003

[4] Cf. Jackson 1977, Tedeschi 1981, Thomason & Gupta 1980, Thomason 1985, Bennett 2003, Ippolito 2003, 2006, 2013b, Arregui 2005b, 2007, 2009, Placek & Müller 2007. Note that such theories need not take any stand on the metaphysics of time. They may understand the history structures as genuine branching or as a bundling of distinct worlds with overlapping pasts.

providing additional support for such theories.

1 Three observations

Let us fix some terminology and draw some observations. Restricting attention to counterfactuals about events for this paper, we distinguish **forward** from **backward** counterfactuals, and distinguish **backtracking** from **non-backtracking** interpretations of the former.[5] Forward counterfactuals are those whose antecedents are about events which take place before the events their consequents are about, while backward counterfactuals are non-forward (those whose antecedents are about events that overlap or take place after the events their consequents are about).[6] (1) is a forward counterfactual that is true only on a non-backtracking interpretation, while (2) is a forward counterfactual true only on a backtracking interpretation. Here is a rough intuitive gloss of the two interpretations. In evaluating a non-backtracking interpretation of a forward counterfactual, one 'punches' its antecedent-event into the causal history of the world and then plays things out from there to see whether its consequent is thereby made true. Such a procedure results in holding fixed the fact that there is nothing underneath Smith during his jump when evaluating (1), which is why it comes out true on its non-backtracking interpretation. In evaluating a backtracking interpretation of a forward counterfactual, by contrast, one does a bit of 'detective work' to figure out in what circumstances its antecedent would have been true, and then, making the requisite changes to history to bring about its antecedent, plays things out accordingly to see whether its consequent is thereby made true. This procedure results in hypothesizing a net being placed underneath Smith to prevent his jump from killing him, which is why (2) comes out true on this backtracking interpretation.

The scenario described at the outset reveals three relevant observations about counterfactuals:

[5] Counterfactuals about states include:

(i) a. If kangaroos had no tails, they would topple over.
b. If the proof had been valid, its premises would have entailed its conclusion.

[6] For now, I will set aside the 'syntactically peculiar' backward counterfactuals which contain an extra 'have to' (as noted by Lewis 1979a), such as:

(i) If Smith had jumped, there would have to have already been a net below him to catch him safely.

I return to discuss these backward counterfactuals in §5.1.

A. Forward counterfactuals admit of two kinds of interpretations: backtracking and non-backtracking.

B. The default interpretation of a forward counterfactual is non-backtracking.

C. Asserting a backward counterfactual will often make salient a backtracking interpretation of a forward counterfactual sharing the same antecedent with, and uttered after, that backward counterfactual.

(A) is well-known and confirmed by the fact that (1) seems to get a non-backtracking interpretation in the original context, while (2) seems to get a backtracking interpretation in the context of Beth's speech. (B) and (C) are new observations. I discuss each in turn.

My evidence for (B) is that, in the context prior to Beth's speech, the most natural interpretation of (1) is non-backtracking—that is why it is most naturally interpreted as true as uttered discourse initially. (B) is also illustrated by the fact that it is (or would be) reasonable and appropriate for you to feel anxiety about Smith's possible jump and relief when Smith does not jump. Notice that, no matter how convincing Beth's story, it is hard to see how it could serve as an argument that your anxiety about the possibility of Smith's jumping and subsequent relief after Smith does not jump are inappropriate or otherwise misguided. Despite the philosophical histrionics Beth pulls, I remain firmly convinced that your relief that Smith did not jump is appropriate, and a reasonable explanation for this is that (1) is true on its default reading, along with the fact that things would have been worse off had Smith died.

Evidence for (C) is that, in setting up the context for her backtracking interpretation of (2), Beth first utters a backward counterfactual ('had Smith jumped, there would already have been a net below him to catch him safely'). Furthermore, this is no quirk about Jackson's example: in every major discussion of backtracking counterfactuals, a backtracking interpretation is brought out in a context in which a similar contextual preamble is asserted, followed by a relevant backward counterfactual, then followed by the target backtracking forward counterfactual. Finally, both (B) and (C) have even been confirmed empirically.[7]

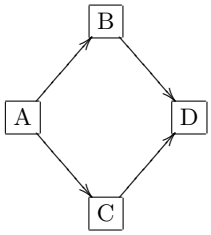
I conclude that any plausible theory of the meaning of counterfactuals ought to predict these three observations about backtracking counterfactuals. However, no existing theory of counterfactuals predicts all three observations. In the next section, I will discuss how

[7]See Gerstenberg *et al.* 2013. In their experiment, they set up as background a causal network as follows (where A causes B and C, and B and C are individually sufficient to cause D):

David Lewis's favored 'similarity theory' of counterfactuals falls short in this respect, and consider the prospects for augmenting the theory to make the right predictions.

2 What similarity lacks

David Lewis once considered adopting a historically structured theory, but rejected it, thinking that its additional constraints rendered it unable to handle certain counterfactuals—for instance those whose antecedents and consequents seemed not about any particular times, as well as backward counterfactuals (cf. Lewis 1979a).[8] Lewis ultimately adopted a less constrained theory on which counterfactuals are about the most similar antecedent-worlds to the evaluation world (see for example Lewis's similarity theory—Lewis 1973a,b, 1979a—as well as Kratzer's lumping semantics—Kratzer 1989, 2012).[9] My goal in this



The design of the experiment involved setting the actual values of A, B, C, and D to *yes* and varying the order of counterfactual questions presented to participants (either starting with the target forward counterfactual or starting with a related backward counterfactual):

(i) Condition 1:
   a. Would D have occurred if B had not occurred?
   b. Would C have occurred if B had not occurred?
   c. Would A have occurred if B had not occurred?

(ii) Condition 2:
   a. Would A have occurred if B had not occurred?
   b. Would C have occurred if B had not occurred?
   c. Would D have occurred if B had not occurred?

Ratings were on a -1 (no), 0 (unsure), 1 (yes) scale. They found that participants in Condition 1 were more likely to say that D would have occurred if B had not ($M = 0.5, SD = 0.82$) than participants in Condition 2 ($M = 0, SD = 0.88$), $t(78) = -2.64, p = .01, d = -0.6$.

[8] We are setting aside counterfactuals apparently about no particular times in what follows here, but, nonetheless, I think Lewis's worries can be met. See Khoo 2015b for a defense of a historically structured theory of counterfactuals against such generality challenges. We will see in §2 that backward counterfactuals are no trouble for my historically structured theory.

[9] Another important class of theories are interventionist theories of counterfactuals (cf. Pearl 2000, Hiddleston 2005, Briggs 2012), which hold that counterfactuals are evaluated relative to causal networks.

section is to provide some reasons for thinking that opting for the less constrained similarity theory is the wrong move. In particular, I will argue that Lewis's theory does not predict (A)–(C), and furthermore that it is not obvious how to amend it so that it does. Since Lewis himself explicitly set aside backtracking counterfactuals, any conclusions of this section are speculative and not decisive. My goal here is just to illustrate the sort of trouble (A)–(C) raise for a standard theory of counterfactuals in order to contrast it with how my theory predicts (A)–(C).

I pause to briefly remark on my choice of terminology for the rest of this paper. Let c be a variable over contexts and w be a variable over worlds. Uppercase italic letters like 'A' denote sentences, '$A \boxrightarrow C$' denotes the English *would*-counterfactual with antecedent A and consequent B, and '$A \diamondrightarrow C$' likewise for the English *might*-counterfactual. Serif uppercase letters like 'A' denote propositions, which I will assume are sets of possible worlds (i.e., subsets of the set of all worlds $\mathcal{W}$). '$\overline{\mathsf{A}}$' denotes the negation of A (set-theoretically, $\mathcal{W} \setminus \mathsf{A}$), '$\mathsf{A} \cap \mathsf{B}$' denotes the conjunction (intersection) of A and C, and so on. Finally, '$\mathsf{A} \models \mathsf{B}$' expresses that A entails B, that is, that every A-world is a B-world.

Following Lewis and many others, I will assume that counterfactuals are context-dependent quantifiers over possible worlds. But, in a given context, what domain of worlds do counterfactuals quantify over? Lewis's proposal is that the domain of a counterfactual $A \boxrightarrow C$ is the set of A-worlds that are most similar (in context c) to w. This yields the following semantics:

> LEWIS:
> $A \boxrightarrow C$ is true at c, w iff all the most similar$_c$ A-worlds to w are C-worlds.

Until we know what the most similar$_c$ worlds are, for a given c, we do not have a predictive theory of the truth conditions of counterfactuals in c. My challenge to LEWIS is to supply an independently motivated account of the contextually supplied similarity relation that, together with his semantics, predicts (A)–(C). Of course, I have no proof that a plausible story here cannot be told on a similarity semantics. My aim in this section is to argue that some reasonable things one might say in response to my challenge do not help. My aim in the rest of the paper is to argue that opting for a historical modality theory *does* help.

However, although sophisticated interventionist theories like that of Hiddleston 2005 do predict (A), even that theory fails to predict (B) and (C) (see especially §4 of Hiddleston 2005). I will not discuss interventionist theories further at this time, except to note that my theory and the interventionist theory make similar predictions about non-backtracking counterfactuals (albeit in slightly different ways); see §2.

Lewis's official semantics for non-backtracking interpretations of counterfactuals comes in Lewis 1979a, where he explicitly sets aside backtracking interpretations to focus on the former, noting that 'only under the standard [non-backtracking] resolution do we have a clear-cut asymmetry of counterfactual dependence that interests me' (p. 458). Lewis focuses on the following objection of Fine 1975 to his earlier proposal that measures worlds by their overall similarity:

> `Nixon`. It is 1975 and President Nixon has just learned that a certain unfriendly country has acquired a nuclear bomb. Before him is the button connected to the arming and firing of several nuclear warheads. He dismisses the idea, instead opting for a strategy of peaceful disarmament.

In this situation, it seems true that:

(3) If Nixon had pressed the button, there would have been a nuclear holocaust.

However, since in fact there was no nuclear holocaust, it seems (by a very natural notion of 'similar') that the most similar worlds in which Nixon presses the button are ones in which no holocaust occurs. Thus, it seems that Lewis's semantics ought to predict that (3) is false—had Nixon pressed the button, the signal would have failed (or something else interfered, preventing the holocaust).

Lewis's response is that such pre-theoretic judgments of similarity are not responsible for our truth conditions of counterfactuals. Rather, a four-part system of weights determines the similarity relation governing non-backtracking counterfactuals (p. 472):

WEIGHTS

1. It is first importance to avoid big, widespread, diverse violations of law.
2. It is of second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
3. It is of 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.

Here is roughly what the analysis predicts for the non-backtracking interpretation of (1), restricting our attention to the case in which the laws are deterministic.

(1) If Smith had jumped, he would have died.

The actual world α is one in which Smith is fully rational, knows there is nothing beneath him to catch him, thus does not jump, and thus does not die. Now, consider four classes of worlds in which Smith jumps:

- w_1 is exactly like α up until some time just before Smith's jump, when some small divergence in α's laws occurs, leading to Smith jumping. At w_1, no net is underneath Smith and hence Smith dies from his jump.

- w_2 contains no divergences from α's laws. Thus, given the assumption that α's laws are deterministic, w_2's history is entirely different from α's. w_2-type worlds will plausibly further subdivide into ones in which there is a net underneath Smith at his jump and ones in which there is no net underneath Smith.

- w_3 is exactly like α up until some time just before Smith's jump, when some small divergence in α's laws occurs, leading to Smith's jump. Then just after Smith jumps, another divergence from α's laws occurs, leading to an outcome similar, though not perfectly similar, to that of α (in particular, Smith does not die, though Smith remembers jumping and so on).

- w_4 is exactly like w_3 except that after Smith jumps a widespread and diverse divergence in α's laws that not only saves Smith but also removes all trace of Smith having jumped, leading to an outcome exactly like that of α.

For LEWIS to predict our intuitions about the non-backtracking interpretation of (1)—in particular, that it is true—it must be the case that w_1 is more similar to α than either w_2, w_3 or w_4. WEIGHTS is designed to predict exactly this fact.[10] Thus, according to WEIGHTS, the most similar worlds in which Smith jumps are like α for most of their history up until Smith jumps; however, each contains a small miracle which leads to Smith deciding to jump (contra his being fully rational, not wanting to die, and knowing there was no net beneath him). Importantly, after this miracle leads to Smith jumping, no other miracle makes a net appear beneath him, and as a result Smith dies from his jump and (1) is true.

[10] Furthermore, WEIGHTS predicts this fact without building an asymmetry between past and future into the similarity relation. This was an important result for Lewis, who was committed to grounding the fact that the future depends asymmetrically on the past in the fact that future events asymmetrically counterfactually depend on past events.

Let us suppose for now that LEWIS + WEIGHTS yields the right predictions regarding the truth conditions of the class of non-backtracking counterfactuals.[11] Can the theory be extended to account for backtracking interpretations of counterfactuals? In the context of Beth's speech, on the backtracking interpretation of (2), all the worlds in its domain in which Smith jumps are worlds in which a net had already been placed beneath him, and hence are all worlds in which he survives the jump. Clearly, WEIGHTS does not yield this prediction. That is fine—Lewis never intended that it do so—but what kind of similarity relation might yield this prediction?

One option is to hold that there are two relevant similarity relations: WEIGHTS (given above), and BACKTRACKING WEIGHTS, which is identical to the former except in that it ranks the importance of `4` above `3`. Thus, according to BACKTRACKING WEIGHTS, securing approximate similarity of particular fact is more important than avoiding small, localized violations of law. As such, by this metric, the most similar worlds in which Smith jumps will be ones which lack any violation of α's laws—therefore, given our assumption that α's laws are deterministic, these will all be worlds which differ from α in matters of fact throughout all of their history. Therefore, this strategy yields 'backtracking unlimited' (Lewis 1979a, p. 171, Bennett 2003, pp. 206–207): the most similar worlds in which Smith jumps will be those in which the entire history of the world is different in certain ways. This clearly will not do for predicting the relevant backtracking interpretation of (2). On that interpretation, (2) comes out true because a net is placed beneath Smith prior to his jump, preventing it from killing him. But, given BACKTRACKING WEIGHTS, the most similar worlds in which Smith jumps will be ones with entirely different histories. And we have no reason think that all such worlds will be ones in which Smith jumps and lives (some may be worlds in which Smith grows up depressed and jumps because he is suicidal, and so on). The lesson from this experiment is that backtracking interpretations of counterfactuals seem to involve *some* backtracking, but not *too much*, and switching the importance of `3` and `4` yields far too much backtracking. Hence, it cannot be what want.

An alternative strategy is to articulate the similarity metric for backtracking interpretations not by modifying some of the principles in WEIGHTS but in some other way.[12] I am

[11] This assumption is itself highly controversial: see Tichý 1976, Slote 1978, Elga 2001, Tooley 2002, 2003, Edgington 2004, Schaffer 2004b, Wasserman 2006. See also Kment 2006, 2014's amended Lewisian theory, which avoids some of these problems. However, though I am sympathetic to some of Kment's ideas (see in particular §3.1), his theory also fails to provide a unified account of backtracking and non-backtracking counterfactuals.

[12] Jonathan Schaffer (personal communication) suggests the following general fix: bump all of the weights

not sure exactly how such a theory will go, but let us suppose one succeeds in doing just this. Still, we will only have a theory that predicts (A). It remains to be seen whether such an emended discussion of the similarity metric will be able to predict (B), which would require explaining why one of the similarity metrics is default, and (C), which would require explaining why uttering a backward counterfactual with the same antecedent as $A \boxdot\!\!\rightarrow C$ is sufficient to make salient the non-default backtracking similarity metric. Again, though I have no proof that such a theory cannot be plausibly motivated, I hope to have shown that quite a bit of work remains for the similarity-theorist.[13] At this point, I think we might be better off exploring whether other, more constrained, theories fare better with respect to (A)–(C). To that end, I turn now to articulating such an alternative theory, one which embraces the historical structural constraints Lewis rejected from the outset.

3 Counterfactuals and histories

As a preliminary motivation for adopting a historically structured semantics for counterfactuals, notice that counterfactuals like (1) are distinguished from their indicative cousins (e.g., (4)) morphologically: the counterfactual contains an extra layer of past tense and

down one step and add a new weight on top: *hold fixed the truth value of all the contextually sacrosanct propositions*. Then, given that in the initial context it is contextually sacrosanct that there is no net underneath Smith, this must also be true at the most similar worlds at which Smith jumps—hence (1) is predicted to be true in the initial context. Somehow, when Beth raises the possibility of Smith's jump being preceded by a net being placed underneath him, this makes that proposition not contextually sacrosanct, and hence allows for the backtracking reading in which (2) comes out false. I think this is a promising proposal, and I cannot fully address it here. However, I will mention three worries. The first is why, in the null (or default) context in which (1) is evaluated, it is contextually sacrosanct that there is no net beneath Smith. Is it because it is mentioned in the description of the case, or because it is visible to all of the parties on the scene? Furthermore, since we know that at least one of the propositions asserted in that preamble must not be held fixed in the evaluation of (1) (given that they are jointly incompatible with the proposition that Smith jumps), unless we have some explanation why the proposition that there is nothing beneath him is different from the others, the 'contextual sacrosanctness' view will not predict (B). Second, I wonder whether and why the proposal that Beth's utterance of the backward counterfactual 'had Smith jumped, there would have been a net underneath him' is sufficient to override the contextual sacrosanctity of there not being a net underneath Smith. Third, I wonder whether, once we have a notion of contextual sacrosanctity in place, we even need Lewis's other weights. If we do not, then this would not amount to an amendment to Lewis but a wholesale replacement of the theory.

[13] Since she does not discuss backtracking counterfactuals, it is unclear what Kratzer's lumping semantics (Kratzer 1989, 2012) would say about (A)–(C). However, notice that what is needed to explain any of these facts via lumping is some context-dependence in what lumps what, or the extent to which lumping determines similarity. Furthermore, it is not obvious how to make either thought precise. Granted, a full examination of the extent to which Kratzer's lumping semantics may be able to explain (A)–(C) lies far beyond this paper, and I will not be able to undertake such an investigation at this time.

the tense auxiliary 'would' in its consequent:[14]

(1) If Smith had jumped, he would have died.

(4) If Smith jumped, he died.

That the extra layer of past tense on (1) is not doing what past tense normally does is illustrated by the fact that the past perfect 'had' felicitously combines with future-oriented frame adverbials like 'tomorrow' in the antecedents of conditionals, but not outside of them:[15]

(5) The contest was held today ...

a. If you had entered tomorrow, you would have missed it.
b. Luckily for Sue, she had (already) entered the contest last night.
c. #Unfortunately for Smith, he had (already) entered the contest tomorrow.

The most straightforward hypothesis about what this extra layer of past tense is doing is shifting the evaluation time of the counterfactual to the past (cf. Tedeschi 1981, Thomason & Gupta 1980, Dudman 1983, 1984, 1988, Edgington 1995, Ippolito 2003, 2006, 2013b, Arregui 2005b, 2007, 2009).[16] Historically structured theories embrace this hypothesis, holding that the past tense allows us to talk about alternate futures which are accessible only from past branch points. I will say more about this in a moment, but we may state the basic idea by appealing to a contextually supplied domain function, D_c, that maps propositions, worlds, and times to sets of worlds:[17]

[14]The past perfect 'had' is normally used to mark that the event described takes place at a time to the past of some reference time, which is itself to the past of speech time (cf. Reichenbach 1947). For instance:

(i) Yesterday, Sue had called the prospects before John got a chance to call them.

Also, it is generally accepted that 'would' is the past of 'will' (cf. Palmer 1986, Ogihara 1996, Abusch 1997, 1998), as in:

(ii) Yesterday, someone would shoot Kennedy.

[15]This feature of subjunctive conditionals has been called 'forward time shift' in the conditionals literature (cf. Gibbard 1981, Dudman 1983, 1984, Edgington 1995, Bennett 2003).

[16]This hypothesis contrasts with one on which the past tense morphology has a 'modal distancing' effect, indicating that the antecedent and consequent may reach beyond the set of worlds that might be actual given what's presupposed in the conversation (the main idea is due to Iatridou 2000, though see also Stalnaker 1975, Isard 1974, Lyons 1977, von Fintel 1997, Starr 2013, Schulz 2014).

[17]Throughout this paper, I will assume for the sake of simplicity that counterfactuals are context-dependent variably strict quantifiers over possible worlds. This assumption is obviously controversial,

SEMANTICS:

$A \mathrel{\Box\!\!\rightarrow} C$ is true at c, w, t iff all worlds in $D_c(\mathsf{A}, w, t')$ are C-worlds. (Where t' is before t)

To fix terminology, call the time that is input to the domain function D *counterfactual time* (Bennett 2003 calls this the time of the fork, or branch).[18] So, what is $D_c(\mathsf{A}, w, t)$? Given our above thought about accessing past-accessible futures, a plausible answer is that $D_c(\mathsf{A}, w, t)$ is some subset of the historically possible A-worlds at w, t (cf. Thomason & Gupta 1980, Thomason 1985, Tedeschi 1981, Ippolito 2003, 2006, Placek & Müller 2007, Arregui 2007, 2009):

HISTORICAL:

For any w, t: $D_c(\mathsf{A}, w, t)$ is a subset of the historically possible A-worlds at w, t.

We define the historically possible worlds relative to w at time t as follows:

Def 1. The **historically possible** worlds relative to world w at time t is the set $H(w, t) = \{w' : w'$ is exactly intrinsically alike w at all times $t' \leq t\}$.

The intuitive appeal of this approach is that historical possibilities are distinguished by being asymmetrically structured with respect to time, so that what was once historically possible may not now be historically possible, while everything that is now historically possible was always historically possible. We can see this visually in the following diagram:

but, fortunately, nothing that I say in the paper turns on these assumptions. We could reformulate our discussion within a strict semantics for counterfactuals, or even within a selection-function semantics in which their truth depends on whether C holds at some particular A-world. For further discussion of such assumptions, see Stalnaker 1968, 1980, Lewis 1973b, von Fintel 2001, Gillies 2007, Swanson 2011.

[18] There will often be differences between nearby times t and t' that don't matter to the truth value of some counterfactual $A \mathrel{\Box\!\!\rightarrow} C$ (holding other things fixed), and in such case a speaker may not intend her utterance of $A \mathrel{\Box\!\!\rightarrow} C$ to have as its counterfactual time t rather than t'. To ensure a common subject matter for a conversation involving some counterfactuals, we should probably not make them about specific times but rather intervals of time (indeed, if time is dense we have no other option anyway). To minimize the complexity in stating the theory, I will ignore this complication in what follows.

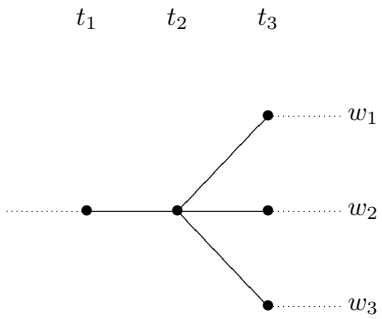


Figure 1

At t_1, w_1, w_2, and w_3 are all historically accessible to w_1, but at t_3 only w_1 is historically accessible to w_1. Supposing that counterfactuals are modals that quantify over historical possibilities offers a plausible explanation for why counterfactuals are past-tensed and why they contain 'would' in their consequents—the past tense allows them to quantify over past historical possibilities that are no longer historically possible, and 'would' situates the time of the consequent to the future of that past history-branch point.

Combining SEMANTICS + HISTORICAL yields that $A \boxdot\!\!\rightarrow C$ is true at c, w, t iff every world in a certain subset of the historically possible A-worlds at w, t' is a C-world (where t' is before t). As it stands, this is just a schema for delivering the truth conditions of counterfactuals—different ways of constraining counterfactual time t' and selecting the relevant subset of the historically possible A-worlds will yield different truth conditions. Thus, the schema is a flexible one, which is good because flexibility seems to be what is required to predict (A), (B), and (C). However, in order for this strategy to work, we need to pair SEMANTICS with the right account of t' and D_c. In the rest of this section, I develop my favored historical-modality theory of counterfactuals, motivating several additional constraints on t' and D_c. In §4, I provide a pragmatic motivation for an additional default constraint on t' and then show how the resulting theory predicts (A)–(C). Each of the constraints discussed are novel to my theory, distinguishing it from other historical-modality theories of counterfactuals.[19]

[19]See for instance Jackson 1977, Tedeschi 1981, Thomason & Gupta 1980, Thomason 1985, Bennett 2003, Ippolito 2003, 2006, 2013b, Arregui 2005b, 2007, 2009, Placek & Müller 2007. None of these authors discuss what factors might set the relevant past counterfactual time. Arregui 2005a discusses backtracking counterfactuals within the context of a historical-modality theory, but comes to very different conclusions from mine. One problem with Arregui's theory is that it is incomplete: her theory only extends to backtracking counterfactuals which contain an extra 'have to'. But clearly there are backwards and backtracking counterfactuals which do not contain the extra morphology (e.g., (2) and (9)).

3.1 Additional constraints on the domain function D_c

Counterfactuals bear a close connection to causation and laws of nature.[20] For instance, as many theorists have noted, law-like and causal statements seem to support counterfactuals:

(6) a. It is a law that water is H_2O.
b. If this substance had been water, it would have been H_2O.

(7) a. John's throwing the rock caused the window to break.
b. If John had not thrown the rock, the window would not have broken.

Furthermore, our intuitions about laws and causation seem to guide our intuitions about counterfactuals. In Fine's Nixon example (reprinted here), we intuitively seem to hold fixed the actual laws in evaluating what happens after the button is pressed.

(3) If Nixon had pressed the button, there would have been a nuclear holocaust.

On such grounds, we might be persuaded to think that $D_c(\mathsf{A}, w, t)$ ought to entail all the laws of nature of w. However, this quickly leads to strange results, if determinism is true. Take an arbitrary counterfactual $A \boxarrow C$. Either its counterfactual time is the first moment of w or not. Supposing its counterfactual time t is the first moment of w, then $A \boxarrow C$ will have an extreme backtracking interpretation in which no facts of w are held fixed (except those which are entailed by the laws and certain settings of the initial conditions). Supposing its counterfactual time t is not the first moment of w, then if $D_c(\mathsf{A}, w, t)$ entails all of w's (deterministic) laws and A is false at w, then $D_c(\mathsf{A}, w, t) = \emptyset$. It thus follows that $A \boxarrow C$ is trivially true. Therefore, if the laws of w are deterministic and $D_c(\mathsf{A}, w, t)$ entails all of them, then either all counterfactuals with false antecedents have extreme backtracking interpretations or are trivially true. This is an undesirable result.

I will not take a stand on determinism or indeterminism in this paper. Nonetheless, I will draw a similar lesson to the one Lewis drew from examples like (3), though I will implement this lesson in a slightly different way (the benefit of this approach will become clear in a moment). Rather than appeal to laws of nature, I will appeal to the notion of a proposition being *causally sufficient* for another (strictly speaking, it is the *truth* of one proposition that is causally sufficient for the *truth* of the other; talking in the former way is a convenient shorthand).[21] Given the notion of causally sufficient, we then define the

[20] See for instance Goodman 1947, Chisholm 1955, Lewis 1986, 2000, Maudlin 2007, Lange 2009.

[21] I will not provide an analysis of what it is for a proposition to be causally sufficient for another. Here

following useful terms:

> **Def 2.** $\mathsf{A} \supset \mathsf{B}$ is a **causal sufficiency** of w iff A is causally sufficient for B at w.
>
> **Def 3.** A is **uniquely causally sufficient** for B at w iff A is causally sufficient for B at w and no other proposition is causally sufficient for B at w.
>
> **Def 4.** $\mathsf{A} \equiv \mathsf{B}$ is a **unique causal sufficiency** of w iff A is uniquely causally sufficient for B at w.

Let $\mathcal{S}_w$ be the set of all causal sufficiencies of w. Finally, let the interval of time a proposition is *about* be the interval of time throughout which the event/state it describes takes place (it need not be continuous). With this in hand, we define the function S which maps a world w and time t to the subset of $\mathcal{S}_w$ about intervals of time entirely to the future of t:

> **Def 5**. $S(w,t) = \{\mathsf{P} \in \mathcal{S}_w : \mathsf{P}$ is about an interval of time that is entirely to the future of $t\}$

$S(w,t)$ are thus the causal sufficiencies of w about times to the future of t. (Notice that, for a causal sufficiency to be about times to the future of t, both its antecedent and consequent must both be about times to the future of t.) From here, we add the following constraint on counterfactual domains:

> CAUSAL:
> $D_c(\mathsf{A}, w, t)$ is a subset of $\bigcap S(w,t)$.

CAUSAL has the effect of constraining $D_c(\mathsf{A}, w, t)$ to contain only worlds which make true every causal sufficiency of w whose antecedent is about times to the future of t. As such, CAUSAL ensures that (3) will come out true, assuming that its counterfactual time is some time after the relevant background facts X are settled, but just prior to Nixon pushing the button. To see why, recall that it is a feature of the case that pushing the button is causally sufficient, given the background facts X and laws, for launching a nuclear missile. So, where $\mathsf{P_t}$ is the proposition that Nixon pressed the button at t and $\mathsf{M_{t'}}$ is the proposition that the nuclear missiles are launched at t', we have that $(\mathsf{P_t} \wedge \mathsf{X}) \supset \mathsf{M_{t'}}$ is a causal sufficiency

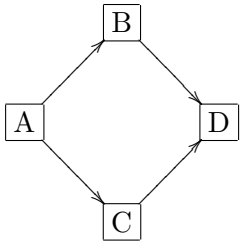
is a (very) minimal constraint: if A is causally sufficient for B then either A is false or B is true. Of course, we do not want the converse to hold, but my hope is that, once we see the role causal sufficiencies play in the theory, we will find them understandable enough to tolerate taking them as primitive for now.

of w.[22] Suppose also that launching nuclear missiles is causally sufficient for there being a nuclear holocaust. Then, given that (3)'s counterfactual time is some time $t^{-1} < t$ at which X is historically necessary but $\mathsf{P_t}$ is not, it follows that:

- $D_c(\mathsf{P_t}, w, t^{-1}) \models \mathsf{X}, \mathsf{P_t}$ By HISTORICAL
- $D_c(\mathsf{P_t}, w, t^{-1}) \models (\mathsf{P_t} \ \wedge \ \mathsf{X}) \supset \mathsf{M_{t'}}$ By CAUSAL

Hence, every world in $D_c(\mathsf{P_t}, w, t^{-1})$ will be one in which $\mathsf{M_{t'}}$ is true, and hence one in which a nuclear holocaust occurs; so (3) will be true. So far, so good.

Nonetheless, HISTORICAL + CAUSAL is not a strong enough constraint on counterfactual domains—it posits a massive break in the causal sufficiencies at counterfactual time, where what we intuitively want is a small, local break. To get a sense of the problem, consider the simple causal model from Gerstenberg *et al.* 2013:



t_1 t_2 t_3

The model supplies us the following causal sufficiencies: $\mathsf{A} \equiv \mathsf{B}, \mathsf{A} \equiv \mathsf{C}, \mathsf{B} \supset \mathsf{D}, \mathsf{C} \supset \mathsf{D}$. Now, consider the counterfactual $\neg B \ \Box\!\!\rightarrow D$. Intuitively, on its non-backtracking interpretation, $\neg B \ \Box\!\!\rightarrow D$ is non-trivially true. However, suppose that its counterfactual time is t_1 (the problem will also arise for earlier choices of counterfactual time; at later times, B will be historically settled, and thus $D_c(\overline{\mathsf{B}}, w, t)$ will be empty and hence $\neg B \ \Box\!\!\rightarrow D$ only trivially true). It follows that:

- $D_c(\overline{\mathsf{B}}, w, t_1) \models \mathsf{A}, \overline{\mathsf{B}}$ By HISTORICAL
- $D_c(\overline{\mathsf{B}}, w, t_1) \models \mathsf{B} \supset \mathsf{D}, \mathsf{C} \supset \mathsf{D}$ By CAUSAL

[22] There is a slight fudge here, because strictly speaking what we need is the background facts to hold at t, not at some time just prior. We ensure that if X holds then $\mathsf{X_t}$ holds as well by way of a hindsight constraint (see below for details).

But notice that, given just these two constraints, $D_c(\overline{\mathsf{B}}, w, t_1)$ does not entail C, and hence does not entail D. Hence, we wrongly predict that $\neg B \boxdot\!\!\rightarrow D$ is false!

Here is why we predict the wrong result. Given just HISTORICAL and CAUSAL, in evaluating a counterfactual $A \boxdot\!\!\rightarrow C$, we consider the class of A-worlds that each shares the history of w up until t, and makes true all of w's causal sufficiencies about times extending beyond t. We thus allow for a break in the causal sufficiencies to make room for the counterfactual's antecedent (as Lewis did), but, incorrectly, we allow for *too much* of a break. Intuitively, the counterfactual's domain must also entail certain facts about times later than its counterfactual time that are not *disrupted* by adding its antecedent to the causal order. Luckily, there is a natural fix that achieves this result, and which also solves an independent problem.

The thought behind the fix follows Jonathan Bennett's 'same causal chain' proposal (see Bennett 2003, pp. 234-237). The basic aim is to ensure that the counterfactual's domain entails every true and salient proposition about times extending beyond t that is caused in the same way at historical antecedent-worlds as it is at w (when relating propositions by 'cause' I mean 'cause to be true'; I will use the simpler phrase throughout). Before we implement the idea formally, let us see how it intuitively helps with the above problem. Recall that HISTORICAL and CAUSAL ensure that $D_c(\overline{\mathsf{B}}, w, t_1) \models \mathsf{A}, \overline{\mathsf{B}}, \mathsf{B} \supset \mathsf{D}, \mathsf{C} \supset \mathsf{D}$; but, intuitively, we also want the counterfactual's domain to entail C and hence (by closure) D. On the proposal floated here, since C is a true and salient (by assumption) proposition about times beyond t_1, we check whether it has the same casual origin at historically accessible $\overline{\mathsf{B}}$-worlds as it does at w. Intuitively, it does: at the historically accessible $\overline{\mathsf{B}}$-worlds (at t_1), C is caused by A, just as at w. Therefore, we restrict the counterfactual's domain so that $D_c(\overline{\mathsf{B}}, w, t) \models \mathsf{C}$ (and thus by closure, D).

To implement this idea a bit more carefully, we first define the **historical/sufficiency antecedent worlds** at w, t as the set of antecedent-worlds that match w's history up until t, and make true all of w's causal sufficiencies about times after t. Formally (suppressing relativization to w):

Def 6. $\mathsf{HS}^{\mathsf{A}}_{\mathsf{t}} = \mathsf{A} \cap H(w,t) \cap \bigcap S(w,t)$ The historical/sufficiency antecedent worlds

We state our third and final constraint on D_c as follows:

HINDSIGHT:
$D_c(\mathsf{A}, w, t)$ entails every true and salient proposition P about times extending beyond

t for which $\mathsf{HS}_t^{\mathsf{A}} \cap \mathsf{P} \neq \emptyset \wedge \forall w' \in \mathsf{HS}_t^{\mathsf{A}} \cap \mathsf{P} : \mathsf{P}$ is caused in the same way at w' as it is at w.

HINDSIGHT formally implements Bennett's proposal: the counterfactual's domain entails all the true and salient post-t propositions which are compatible with $\mathsf{HS}_t^{\mathsf{A}}$ and which are caused in the same way at those worlds as at w.

I turn now to argue that this fix is not ad hoc, but in fact allows the theory to handle a well-known problem case in the literature. Consider the following situation (cf. Slote 1978, Barker 1998, Bennett 2003, Edgington 2004, Schaffer 2004a, Noordhof 2005, Kaufmann 2005, Phillips 2007, 2011, Walters 2009, Won 2009, Ahmed 2010, 2011, Arregui 2009, Ippolito 2013b,a): Joe pushes the button on an indeterministic coin-flipping machine which initiates a coin flip; as Joe pushes the button, Sue bets that it will land heads; the coin lands tails and Sue loses the bet. Intuitively, in this context (8-a) is true and (8-b) is false (or at least not true):

(8) a. If Sue had bet on tails, she would have won. True

b. If Sue had pushed the button, the coin would have landed tails. False / indeterminate

The difference between (8-a) and (8-b) seems to result from holding fixed that the coin lands tails in evaluating (8-a) but not (8-b). HINDSIGHT allows my theory to do just this. Let t be some time just before Sue does not bet on tails, $\mathsf{B_T}$ be the proposition that Sue bet on tails, and T be the proposition that the coin landed tails. Consider (8-a) first. By HISTORICAL and CAUSAL, $\mathsf{HS}_t^{\mathsf{B_T}}$ will contain only worlds just like w up until t at which Sue bets on tails, and which are causally alike w after t. T is caused in the same way at worlds in $\mathsf{HS}_t^{\mathsf{B_T}}$ as it is at w, at both such worlds, T is caused by Joe pushing the button.[23] Thus, presuming for now that T is contextually salient, then by HINDSIGHT, $D_c(\mathsf{B_T}, w, t)$ will entail T and hence entail that Sue won.

Compare this result with (8-b), and let t' now be some time just before Joe pushes the button. Let $\mathsf{P_S}$ be the proposition that Sue pushes the button. By HISTORICAL and CAUSAL, $\mathsf{HS}_{t'}^{\mathsf{P_S}}$ will contain only worlds just like w up until t' at which Sue pushes the button (instead of Joe), and which are causally alike w after t'. Now, notice that T is not caused in the same way at worlds in $\mathsf{HS}_{t'}^{\mathsf{P_S}}$ as it is at w: at w, T is caused by Joe pushing

[23] This is one of the payoffs of stating the semantics in terms of causal sufficiencies rather than laws: we need not presume that every causal sufficiency is an instance of some deterministic law. Thus, we allow for indeterministic causation, as seems to be the case in this coin flip example.

the button, whereas at worlds in $\mathsf{HS}^{\mathsf{P_S}}_{t'}$, T is caused by Sue pushing the button. Therefore, given that HINDSIGHT is our final constraint on counterfactual domains, $D_c(\mathsf{P_S}, w, t')$ will not entail T, and hence we predict that (8-b) is false.[24] I conclude that HINDSIGHT is an independently motivated constraint on counterfactual domains.

To sum up, the domain of a counterfactual $A \boxdot\!\!\rightarrow C$ at c, w, t is $D_c(\mathsf{A}, w, t)$. This is the set of A-worlds that match the history of w up until some time t (counterfactual time), match w's causal sufficiencies thereafter, and entail all the c-salient post-t facts that are not 'disrupted' by supposing A. We turn next to motivating some minimal constraints on counterfactual time, t.

3.2 Minimal constraints on counterfactual time

What sort of minimal constraint might we motivate for counterfactual time? A plausible candidate is:

> **Latest antecedent support**: For any counterfactual $A \boxdot\!\!\rightarrow C$, its counterfactual time t is the latest past time such that $D_c(\mathsf{A}, w, t) \neq \emptyset$.

Latest antecedent support states that the counterfactual time for any counterfactual $A \boxdot\!\!\rightarrow C$ is the latest past time at which it has a nonempty domain. Combining **Latest antecedent support** with HISTORICAL, CAUSAL, and HINDSIGHT yields a semantics for counterfactuals that is equivalent (as far as I can tell) to that endorsed by Jonathan Bennett (see Bennett 2003, pp. 209–220). Call this theory JBAT, which stands for JUST BEFORE ANTECEDENT TIME:

> JBAT: $A \boxdot\!\!\rightarrow C$ is true at c, w, t iff all worlds in $D_c(\mathsf{A}, w, t')$ are C-worlds.
>
> (Where t' is the latest time to the past of t such that $D_c(\mathsf{A}, w, t') \neq \emptyset$)

[24] The theory can also predict the falsity (or at least non-truth) of (i) in the scenario above:

(i) If Sue had pushed the button, she would have lost.

The proposition that Sue lost, L, is caused by the coin landing tails and Sue betting heads at w. But there are L-worlds in $\mathsf{HS}^{\mathsf{P_S}}_{t'}$ in which Sue bet tails and the coin landed heads. Hence, L is not caused in the same way at w as it is at $\mathsf{HS}^{\mathsf{P_S}}_{t'}$. Furthermore, this holds even if we change the case slightly so that Sue makes her bet before the coin is flipped, so that $\mathsf{HS}^{\mathsf{P_S}}_{t'} \models \mathsf{B_H}$. In that case, L *still* is not caused in the same way at w as it is at worlds in $\mathsf{HS}^{\mathsf{P_S}}_{t'}$. At w, L is caused by Sue betting heads and Joe flipping the coin (causing it to land tails). But at L-worlds in $\mathsf{HS}^{\mathsf{P_S}}_{t'}$, L is caused by Sue betting heads and Sue flipping the coin (causing it to land tails).

This is not a bad first pass at a semantics for counterfactuals. However, **Latest antecedent support** suffers from at least two problems: (i) it cannot handle backward counterfactuals, and (ii) it is unmotivated (as it stands). Focus on (i) for now. I will argue that backward counterfactuals demand a looser minimal constraint on t. To see why, consider the backward counterfactual that Beth utters:

(9) Had Smith jumped, there would have already been a net below him to catch him safely.

Since in fact there was no net beneath Smith, the consequent of (9) is actually false. Let J be the proposition that Smith jumped, and N be the proposition that there was a net beneath him just prior. Therefore, presuming that the latest time t at which $D_c(\mathsf{J}, w, t)$ is non-empty is after the time at which N is about,[25] JBAT predicts that (9) should be false, simply because N is actually false. But this is wrong—the truth of (9) is not determined entirely by whether its consequent is actually true or false. Rather, whether (9) is true or false depends on what would have preceded Smith's jump—and this fact is independent of whether that event actually occurred. The problem here is with **Latest antecedent support**, and can be seen clearly in the following diagram:

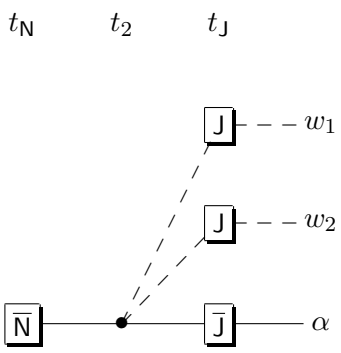


Figure 2

In the diagram, suppose t_2 is the latest time for which $D_c(\mathsf{J}, \alpha, t_2)$ is non-empty. Then, since N is actually false and about times before t_2, it must be false at all historically possible worlds at α, t_2, and hence false at every world in $D_c(\mathsf{J}, \alpha, t_2)$. In such conditions, JBAT predicts (9) to be false. However, intuitively, the truth values of many backward

[25] I take it that this assumption is plausible for (9), though strictly speaking we just need it to be true for at least some backward counterfactuals.

counterfactuals (including (9)) are independent of the truth values of their consequents. Hence, **Latest antecedent support** is too strict—it incorrectly predicts that the truth values of such backward counterfactuals are determined entirely by the truth values of their consequents.

3.2.1 Motivating two constraints

The most natural fix is to reject **Latest antecedent support** in favor of a looser minimal constraint on counterfactual time which allows us to avoid predicting that $A \boxdot\!\!\rightarrow C$ is true/false merely because its consequent is true/false. I propose that following two minimal constraints on counterfactual time:[26]

> The default counterfactual time t for $A \boxdot\!\!\rightarrow C$ will be some time to the past of the time of the context such that (if any):
>
> **Possible antecedent**: $D_c(\mathsf{A}, w, t) \neq \emptyset$.
>
> **Contingent consequent**: $H(w,t) \cap \mathsf{A} \cap \mathsf{C} \neq \emptyset$ and $H(w,t) \cap \mathsf{A} \cap \overline{\mathsf{C}} \neq \emptyset$

Possible antecedent says that the default counterfactual time for some counterfactual $A \boxdot\!\!\rightarrow C$ is one at which its domain is non-empty. **Contingent consequent** says that this time is one at which C is not settled by the history up until t, together with A (neither it nor its negation are A-historically necessary). In the rest of this section, I will motivate both constraints, and then show how they mark an improvement over **Latest antecedent support**.[27]

[26] This proposal may face challenges from wild counterfactuals whose antecedents and consequents are always historically impossible. I want to set aside these challenging counterfactuals for now. There are several ways of handling them—extending the domain of counterfactuals to include impossible worlds, for instance. See Khoo 2015b for more discussion of several strategies.

[27] I pause to emphasize that both are *default* constraints, and hence compatible with exceptions. For instance, we sometimes use counterfactuals in stating reductio ad absurdum reasoning as applied to some hypotheses under consideration. Thus, we have:

(i) A: If Smith had jumped, there would have been a net beneath him to catch him.
B: I agree, but it simply was not possible for there to have been a net beneath Smith to catch him.

If you are like me, you might initially find B's response somewhat strange. This is further support for **Contingent consequent**. However, notice that we can make sense of what B is saying here. In effect, B is trying to reason to the conclusion that Smith did not jump. My theory can predict this if supplemented with a suitable semantic theory of metaphysical possibility modals. I will not defend any particular theory

Begin with **Possible antecedent**. It is plausible that ordinary speakers and hearers aim to avoid saying, and interpret each other as having said, propositions which are trivially true—such propositions are uninformative, and hence of no interest in conversations where information exchange is an important goal. Indeed, this general tendency to avoid trafficking in trivially true propositions seems to be why universal quantifiers invariably carry presuppositions that their domains are nonempty (cf. Strawson 1952, Hart 1951).[28] Here is an example involving nominal quantifiers:

(10) a. #Every living Civil War veteran attended last night's gala.
b. #Did every living Civil War veteran attend last night's gala?
c. #Not every living Civil War veteran attended last night's gala.

Suppose the domain of the restricted nominal quantifier 'every$_\phi$' is the set of things that are ϕ. But the set of living Civil War veterans (the denotation of 'living Civil War veteran') is empty since there are no such things. Thus, the domain of 'every living Civil War veteran' is empty. So if each sentence in (10) presupposes that its primary quantifier has a nonempty domain, each sentence in (10) suffers from presupposition failure, which explains why these sentences are infelicitous. Exactly the same behavior is exhibited by adverbial quantifiers:

(11) a. #Michael Jordan always brought cookies when he went to the moon.
b. #Did Michael Jordan always bring cookies when he went to the moon?
c. #Michael Jordan did not always bring cookies when he went to the moon.

Suppose the domain of the restricted adverbial quantifier 'always$_\phi$' is the set of actual situations that are ϕ (cf. Berman 1987, Kratzer 1989, von Fintel 2004). There are no

at this time, but consider the proposal of Williamson 2007:

(ii) $\Diamond A$ is true at c, w, t iff $A \boxdot\!\!\!\rightarrow \bot$ is false at c, w, t.

Then, supposing each is evaluated relative to the same past time t, if $A \boxdot\!\!\!\rightarrow C$ and $\neg\Diamond C$ are both true, it follows that $\neg\Diamond A$ is also true. (Proof: suppose $A \boxdot\!\!\!\rightarrow C$ and $\neg\Diamond C$ are both true at c, w, t. Then $D_c(\mathsf{A}, w, t) \subseteq \mathsf{C}$ and $D_c(\mathsf{C}, w, t) \subseteq \bot$. Thus, $D_c(\mathsf{C}, w, t) = \emptyset$. Hence, $\mathsf{HS}_t^{\mathsf{C}} = \emptyset$; this follows because it cannot be the case that $D_c(\mathsf{C}, w, t)$ is empty due to it entailing some post-t propositions incompatible with C, since by HINDSIGHT any such propositions it entails *must* be compatible with C. Suppose for reductio that $D_c(\mathsf{A}, w, t) \neq \emptyset$. Then, since $D_c(\mathsf{A}, w, t) \subseteq \mathsf{C}$ (by assumption), it must be that $\mathsf{HS}_t^{\mathsf{A}} \cap \mathsf{C} \neq \emptyset$. But we have already seen that $\mathsf{HS}_t^{\mathsf{C}} = \emptyset$. So, $D_c(\mathsf{A}, w, t) = \emptyset$. Thus $D_c(\mathsf{A}, w, t) \subseteq \bot$, and hence $A \boxdot\!\!\!\rightarrow \bot$ is true at c, w, t. And thus $\neg\Diamond A$ is true at c, w, t.)

[28] I will remain neutral with respect to whether the *sentence* presupposes this, or whether *speakers* using such sentences presuppose this, or whether both do and that there's some explanation of the one in terms of the other.

actual situations in which Michael Jordan went to the moon. Thus, the domain of the restricted quantifier 'always ψ when Michael Jordan went to the moon' is empty. So if each sentence in (11) presupposes that its primary quantifier has a nonempty domain, each sentence in (11) suffers from presupposition failure, which explains why these sentences are infelicitous.

The presumption against triviality seems to be quite robust and general across universal quantifiers of different types. Thus, there is every reason to suppose that there is a similar presumption against triviality in modal expressions, such as counterfactuals. Next, given our semantics, counterfactuals are equivalent to restricted universal quantifiers. Therefore, we have every reason to expect that counterfactuals too will carry a presupposition that their domains are nonempty:

EXISTENCE: $A \boxdot\!\!\rightarrow C$ presupposes that $D_c(\mathsf{A}, w, t) \neq \emptyset$.

Now, recall that, on my theory, the lexical domain of a counterfactual is determined by an independent parameter of interpretation—counterfactual time. Thus, given the default pressure to interpret sentences so that their presuppositions are met, we should expect that the default interpretation of a counterfactual will be one on which its counterfactual time ensures its domain is non-empty, and this is just **Possible antecedent**.

Turn next **Contingent consequent**. Suppose $A \boxdot\!\!\rightarrow C$ is interpreted as per the default (and hence has a non-empty domain). Then, $A \boxdot\!\!\rightarrow C$ is true only if there are some C-worlds in $D_c(\mathsf{A}, w, t)$ (where t is its counterfactual time), and this is so only if $H(w, t) \cap \mathsf{A} \cap \mathsf{C} \neq \emptyset$ (if C is false at w). Now, it seems plausible that considerations of charity will generally motivate interpreting sentences (or utterances thereof) so they have some chance of being true.[29] Thus, we expect the default interpretation of $A \boxdot\!\!\rightarrow C$ to be one in which its counterfactual time allows some C-worlds in its domain (if there are any such times); to fail to interpret $A \boxdot\!\!\rightarrow C$ in this way would be to interpret it in such a way that it has no chance at being true. Admittedly, this is only evidence for one half of **Contingent consequent**. What about its second conjunct, that $H(w, t) \cap \mathsf{A} \cap \overline{\mathsf{C}} \neq \emptyset$?

One motivation for the second conjunct of **Contingent consequent** has to do with counterfactuals with true consequents, such as:

(12) If Smith had jumped, there (still) would have been no net beneath him.

[29] This is a corollary of my principle TRUTH, which I will discuss in more detail in §4.

Although discussing the correct semantics for such 'concessive' counterfactuals goes beyond the scope of this paper,[30] it seems that such counterfactuals place the same restriction on their counterfactual times as counterfactuals with false consequents—in particular, that their counterfactual times are constrained to be prior to the historical settling of their consequents. Without such a restriction—for instance, if we just had the first conjunct of **Contingent consequent**, (12) could have any time as its counterfactual time. But if that were so, combining this proposal with considerations of charity in speaker interpretation (see §4.2 below) will yield a theory which predicts that (12) will automatically be judged true (since there is an easy available interpretation on which it is true: just let its counterfactual time be some time after which it is settled that there was no net beneath Smith). Furthermore, this prediction will apply equally to any backward counterfactual with a true consequent, resulting in the trivialization of all such counterfactuals. But intuitively not all such counterfactuals are trivialized in this way. For instance, in certain contexts, (12) is judged true. But in other contexts (such as after Beth's preamble, as we saw at the outset), it seems false and instead (9) seems true.

(9) Had Smith jumped, there would have already been a net below him to catch him safely.

Thus, backward counterfactuals with true consequents are not automatically judged true. The second conjunct of **Contingent consequent** allows us to avoid this consequence by constraining counterfactual time to be some time before C is historically settled.[31]

Turn next to the upshot of these default constraints on counterfactual time for our historical-modality theory. Suppose A and C are both false at w, as is the case with most counterfactuals. By **Possible antecedent**, the default counterfactual time for $A \boxdot\!\!\rightarrow C$

[30] For instance, the proper account will need to say something about the semantics of words like 'even' and 'still'. See Bennett 1982, 2003, Barker 1991, 1994, Lycan 1991, 2001 for discussion.

[31] I pause to briefly note two other motivations for **Contingent consequent**. Khoo 2015b shows that **Contingent consequent** follows from a principle that, together with a historical-modality theory of counterfactuals, predicts the semantic difference between indicative and subjunctive conditionals like the following:

(i) a. If Oswald did not shoot Kennedy, someone else did.
b. If Oswald had not shot Kennedy, someone else would have.

I will not be able to discuss this additional motivation for **Contingent consequent** in this paper for sake of space. However, I do want to flag that we will see one other important benefit of **Contingent consequent** in §4.2.2. Hence, at least within the context of my overall theory, **Contingent consequent** seems strongly supported.

will be some time before the time A is about.[32] This seems right: remember that, in predicting the right truth values for the counterfactuals in the last section, I assumed that their counterfactual times preceded the times their antecedents were about. For analogous reasons, by **Contingent consequent**, the default counterfactual time for $A \boxdot\!\!\rightarrow C$ will be some time before the time C is about. This is what allows for the possibility of backward counterfactuals whose truth values are independent of their consequents' truth values. In the case illustrated by Figure 2, we suppose that t_1 is the latest moment such that $H(w, t_1) \cap \mathsf{A} \cap \mathsf{N} \neq \emptyset$ and $H(w, t_1) \cap \mathsf{A} \cap \overline{\mathsf{N}} \neq \emptyset$. By **Contingent consequent**, counterfactual time t must not come after t_1. Supposing that its counterfactual time just is t_1, then the truth or falsity of a backward counterfactual (9) ($J \boxdot\!\!\rightarrow N$) will not be settled by the truth value of its consequent. For instance, Figure 3 reveals a scenario in which N is actually false, but $J \boxdot\!\!\rightarrow N$ is actually true:

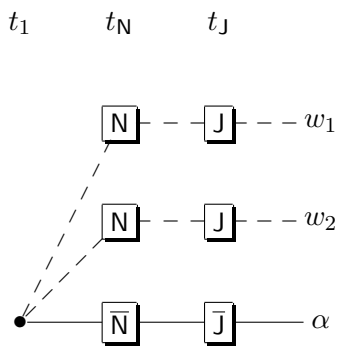


Figure 3

Thus, the two minimal constraints **Possible antecedent** and **Contingent consequent** on counterfactual time are both independently motivated and allow for the possibility of backward counterfactuals whose truth values are independent of their consequents' truth values. These two principles together comprise the default minimal constraints on the counterfactual time of *any* counterfactual (forward, backward, backtracking, non-backtracking). Say that a time t is a *default admissible* counterfactual time for $A \boxdot\!\!\rightarrow C$ in context c iff t satisfies both **Possible antecedent** and **Contingent consequent**.

In the next section, I motivate two general pragmatic constraints on interpretation and show how they affect the choice of counterfactual time. I then show how the resulting

[32] Suppose for reductio that t is some time after the time A is about. Then, given that A is false at w, there are no historically possible A-worlds at t. Thus, by HISTORICAL, $D_c(\mathsf{A}, w, t) = \emptyset$. Hence, **Possible antecedent** is violated.

semantics predicts (A)–(C).

4 Backtracking and time

Recall (A)–(C):

A. Forward counterfactuals admit of two kinds of interpretations: backtracking and non-backtracking.

B. The default interpretation of a forward counterfactual is non-backtracking.

C. Asserting a backward counterfactual will often make salient a backtracking interpretation of a forward counterfactual sharing the same antecedent with, and uttered after, that backward counterfactual.

My strategy is to explain (A)–(C) by showing how certain independently motivated pragmatic principles of interpretation yield choices of counterfactual times that predict (A)–(C). To get a sense of the strategy, consider the theory so far. On the theory, the truth of $A \boxdot\!\!\rightarrow C$ depends on the history of w up until t, the causal sufficiencies of w after t, and the post-t facts not disrupted by A. Earlier counterfactual times hold fixed less of w's past and more of its causal sufficiencies, while later counterfactual times hold fixed more of w's past and fewer causal sufficiencies. Hence, we can think of the choice of counterfactual time t as trading off history for causal sufficiencies (opting for an earlier time) or vice versa (opting for a later time), as bounded by our default admissibility constraints on t, **Contingent consequent** and **Possible antecedent**. My strategy is to connect later admissible counterfactual times to non-backtracking interpretations and earlier admissible counterfactual times to backtracking interpretations. I will argue that there is a default preference for later admissible counterfactual times, thus predicting (B); however, accepting a (relevant) backward counterfactual can override this default and lead to an earlier counterfactual time, thus resulting in a backtracking interpretation—hence predicting (C) and thereby also (A).

Before we get into the theory, it will be helpful to have an idealized picture of the causal situation described at the outset (Smith poised to jump on the roof of the twenty story building):

Facts:

- B: Smith believed that jumping would kill him.
- D: Smith desired to live.
- R: Smith was rational.
- $\overline{\mathsf{N}_{\mathsf{t}_\mathsf{i}}}$: there was not net underneath Smith to break his fall at t_i.
- K: Smith knew whether there was something beneath him to break his fall at t.
- J: Smith jumped at t'.
- L: Smith lived.
- S: If there is no net underneath Smith to break his fall and he knows whether there is, he will believe that jumping will kill him.

  $(\overline{\mathsf{N}_\mathsf{t}} \wedge \mathsf{K}) \supset \mathsf{B}$

- L_1: Anyone who is rational, has no wish to die, and believes ϕing will kill them, will not ϕ.

  $(\mathsf{B} \wedge \mathsf{D} \wedge \mathsf{R}) \supset \overline{\mathsf{J}}$ (Relevant instance)

- L_2: Anyone jumping from a 20 story building with no net underneath them will die.

  $(\mathsf{J} \wedge \overline{\mathsf{N}_{\mathsf{t}'}}) \supset \overline{\mathsf{L}}$ (Relevant instance)

Given this idealization, we may represent the causal structure of the scenario in the following diagram, which represents the causal relations among different propositions at the world of evaluation:[33]

[33] Of course, the time of Smith's hypothetical jump is underspecified by (1). However, it may still be constrained in certain contexts. For instance, supposing that Smith was on the roof from 1:00pm to 1:30pm, then any hypothetical jump of Smith occurring in that interval would count as a relevant hypothetical jump. More on how my theory can be extended to handle underspecified antecedents in §5.5.

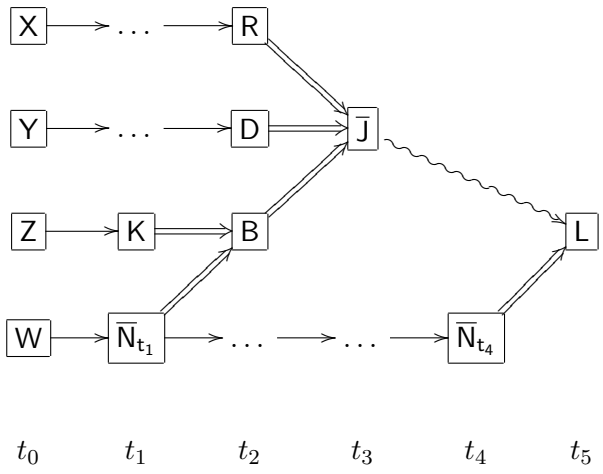
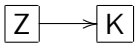
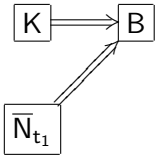


Figure 4

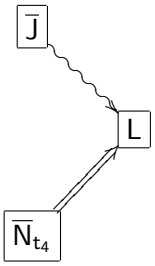
Here is how to read this diagram. Propositions in a box, like $\boxed{\mathsf{R}}$, are true at the evaluation world w. As before, a single arrow $\rightarrow$ between two propositions signifies causal sufficiency—so, for instance,



encodes that $\mathsf{Z} \supset \mathsf{K}$ is a causal sufficiency. A double arrow $\Rightarrow$ signifies joint causal sufficiency—so, for instance,



encodes that $(\mathsf{K} \wedge \mathsf{N}_{t_1}) \supset \mathsf{B}$ is a causal sufficiency. Finally, a squiggly arrow $\rightsquigarrow$ signifies a 'masking condition' such that,



encodes that $(\mathsf{J} \wedge \overline{\mathsf{N}}_{\mathsf{t}_4}) \supset \overline{\mathsf{L}}$ is a causal sufficiency. I suppose that $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$, and W are (suppressed) causal antecedents of $\mathsf{R}, \mathsf{D}, \mathsf{K}$, and $\overline{\mathsf{N}}_{\mathsf{t}_1}$ respectively. Finally, I am taking liberties with the times to keep things simple. There is very likely a large gap between t_0 and t_1 and a much smaller gap between t_2 and t_3.

4.1 Backtracking/non-backtracking and counterfactual time

The first order of business is to show how my theory predicts backtracking/non-backtracking interpretations as arising from the choice of later/earlier counterfactual times. In §4.2, I will argue that, by default, counterfactuals are interpreted as having later counterfactual times, and hence by default have non-backtracking interpretations. In §4.3, I will discuss how we can overrule this default and thus allow for backtracking interpretations.

Consider (1) again:

(1) If Smith had jumped, he would have died.

I show here why the choice of counterfactual time is crucial to whether (1) comes out true on my semantics. Notice first of all that t_0, t_1, and t_2 are all default admissible counterfactual times for (1), but later times are not. I will show how my semantics predicts different interpretations of (1) depending on which counterfactual time it receives, with later counterfactual times generating the (intuitively true) non-backtracking interpretation, and earlier counterfactual times generating the (intuitively true) backtracking interpretation of (2) (in the context of Beth's speech).

First, let (1)'s counterfactual time be the latest default admissible time t_2. Then $\mathsf{X}, \mathsf{Y}, \mathsf{Z}, \mathsf{W}, \mathsf{K}, \mathsf{R}, \mathsf{D}, \mathsf{B}$, $\overline{\mathsf{N}}_{\mathsf{t}_1}$, and $\overline{\mathsf{N}}_{\mathsf{t}_2}$ are all historically settled, and hence entailed by $\mathsf{HS}^{\mathsf{J}}_{\mathsf{t}_2}$. Furthermore, notice that $\overline{\mathsf{N}}_{\mathsf{t}_3}$ is caused in the same way at worlds in $\mathsf{HS}^{\mathsf{J}}_{\mathsf{t}_2}$ as it is at w, being caused by $\overline{\mathsf{N}}_{\mathsf{t}_2}$. Hence, supposing that it is salient that there is no net underneath Smith at t_3 (the time of his hypothetical jump), by HINDSIGHT, $D_c(\mathsf{J}, w, t_2) \models \overline{\mathsf{N}}_{\mathsf{t}_3}$. By CAUSAL, $D_c(\mathsf{J}, w, t_2) \models (\overline{\mathsf{N}}_{\mathsf{t}_3} \supset \overline{\mathsf{N}}_{\mathsf{t}_4}), (\mathsf{J} \wedge \overline{\mathsf{N}}_{\mathsf{t}_4}) \supset \overline{\mathsf{L}})$. Thus, by closure, $D_c(\mathsf{J}, w, t_2) \models \overline{\mathsf{N}}_{\mathsf{t}_4}, \overline{\mathsf{L}}$. Therefore, if (1) has t_2 as its counterfactual time, my semantics predicts that it is true. This is its non-backtracking interpretation, on my semantics.[34]

Next, suppose (1)'s counterfactual time is t_0. Then $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$, and W are historically settled and hence entailed by $\mathsf{HS}^{\mathsf{J}}_{\mathsf{t}_0}$. Furthermore, each of $\mathsf{R}, \mathsf{D}, \mathsf{K}$, and $\overline{\mathsf{N}}_{\mathsf{t}_1}$ are caused in the

[34] Crucially, the proposition that Smith lived, L, is not entailed by $D_c(\mathsf{J}, w, t_2)$. This is because L is not caused in the same way at L-worlds in $\mathsf{HS}^{\mathsf{J}}_{\mathsf{t}_2}$ as it is at w. At w, L is caused by Smith's not jumping. At L-worlds in $\mathsf{HS}^{\mathsf{J}}_{\mathsf{t}_2}$, L is caused by Smith jumping and landing safely on a net.

same way at the worlds in $\mathsf{HS}^{\mathsf{J}}_{t_0}$ where they are true as at w. But then, given the causal sufficiencies of w after t_0, $D_c(\mathsf{J}, w, t_0)$ will entail $\mathsf{R}, \mathsf{D}, \mathsf{K}$, and $\overline{\mathsf{N}}_{t_1}$ only if it is empty.[35] Therefore, at least one of $\mathsf{R}, \mathsf{D}, \mathsf{K}$, or $\overline{\mathsf{N}}_{t_1}$ must not be salient in the context. Furthermore, the choice of which to give up in evaluating the counterfactual affects whether (1) is true, given t_0 as its counterfactual time. The main point here is that if (1) receives t_0 as its counterfactual time, it has a backtracking interpretation on my theory—and depending on which propositions are salient, on some backtracking interpretations it is false.

The point of this exercise is to illustrate how the choice of counterfactual time affects the interpretation of a counterfactual on my semantics, which shows how my semantics can predict (A). Before we get to the details of how my theory predicts (B) and (C), I want to demonstrate (i) how my theory predicts the truth of (9), which is the counterfactual Beth utters to set up her intended backtracking interpretation of (2), and (ii) how my theory predicts the truth of (2), assuming that it receives the same counterfactual time as (9) and is interpreted with respect to the same salient propositions:

> 'Smith was rational, had no wish to die, could see below him, and knew that jumping without a net would kill him. Therefore,
>
> (9) Had Smith jumped, there would have already been a net below him to catch him safely.
>
> So,
>
> (2) If Smith had jumped, he would have lived.'

(i) Predicting the truth of (9)

First off, notice that, by **Contingent consequent**, the default of interpretation of (9) will be one on which it has t_0 (or some earlier time) as its counterfactual time. Second, notice that, in her preamble, Beth asserts R, D, and K. Very naturally, in doing so, she makes each contextually salient (more on salience in §4.3). Therefore, since it is not made salient by Beth's speech, most naturally, $\overline{\mathsf{N}}_{t_1}$ is not one of the salient post-t facts. But, by HINDSIGHT, $D_c(\mathsf{J}, w, t_0) \models \mathsf{R}, \mathsf{D}, \mathsf{K}$, and by CAUSAL, $D_c(\mathsf{J}, w, t_0) \models (\mathsf{K} \wedge \overline{\mathsf{N}}_{t_1}) \supset \mathsf{B}, (\mathsf{R} \wedge \mathsf{D} \wedge \mathsf{B}) \supset \overline{\mathsf{J}}, \mathsf{N}_{t_1} \supset \mathsf{N}_{t_4}$. Therefore, by closure, $D_c(\mathsf{J}, w, t_0) \models \overline{\mathsf{B}}, \mathsf{N}_{t_1}, \mathsf{N}_{t_4}$.

[35] Quick proof. Suppose that $D_c(\mathsf{J}, w, t_0) \models \mathsf{R}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{t_1}$. By CAUSAL, $D_c(\mathsf{J}, w, t_0) \models (\mathsf{K} \wedge \overline{\mathsf{N}}_{t_1}) \supset \mathsf{B}, (\mathsf{R} \wedge \mathsf{D} \wedge \mathsf{B}) \supset \overline{\mathsf{J}}$. Then, by closure $D_c(\mathsf{J}, w, t_0) \models \mathsf{J}, \overline{\mathsf{J}}$.

Thus, we predict the truth of (9) given t_0 as its counterfactual time.

(ii) Predicting the truth of (2), assuming it and (9) are likewise interpreted
We assume that (2) has as its counterfactual time t_0, the same as (9), and also that (2) is interpreted with respect to the same salient propositions $\mathsf{R}, \mathsf{D}, \mathsf{K}$. From (i), we know that $D_c(\mathsf{J}, w, t_0) \models \mathsf{N}_{\mathsf{t}_1}, \mathsf{N}_{\mathsf{t}_4}$. But then, given that the net is a safe one, an additional causal sufficiency of w is $(\mathsf{J} \wedge \mathsf{N}_{\mathsf{t}_4}) \supset \mathsf{L}$. Hence, by CAUSAL, $D_c(\mathsf{J}, w, t_0) \models (\mathsf{J} \wedge \mathsf{N}_{\mathsf{t}_4}) \supset \mathsf{L}$, and thus it follows that $D_c(\mathsf{J}, w, t_0) \models \mathsf{L}$. Therefore, we predict that (2) is true, as long as it has the same counterfactual time and is interpreted with respect to the same salient propositions as (9).

Thus, my semantics is flexible enough to predict a true non-backtracking interpretation of (1), and a true backtracking interpretation of (2). However, we have not yet seen why the non-backtracking interpretation is the default, nor have we seen why (2) should have the same counterfactual time and be interpreted with respect to the same salient propositions as (9). I will address the first topic in §4.2, and explore the latter in §4.3.

Before we move on, I want to illustrate one interesting upshot of my strategy, which is that it predicts that the backtracking/non-backtracking distinction is not a binary one. Instead, it predicts that there are more or less backtracking interpretations, depending on the relative lateness of the counterfactual's time. This might seem surprising, given that we started with a binary distinction, but a bit of reflection reveals that our starting examples are really limiting cases of a more general phenomenon. Here is an example inspired by Lewis 1973a:

(13)
A: Smith was friends with everyone at the party. If he had gone, he would have had fun!
B: But Smith would have gone only if Sue had too, and the only way Sue would have gone is if it had been an 80s-themed party, and Smith hates 80s-themed parties. So, if he had gone, he would not have had fun.
A: I grant that, but Smith would have known about the theme, and hence had he gone, he would not have hated 80s themed parties. So, if he had gone, he would have had fun.

Once we get to A's second utterance of 'if he had gone, he would have had fun' it is

clear that what A means by it is something very different from what A meant by her first utterance of it, even though both are true! In evaluating A's first counterfactual, we suppose things are pretty much as they are except that Smith goes to the party. In evaluating B's counterfactual, we backtrack to make room for Smith going and this requires supposing that it would also have been an 80s party. But then when we get to A's second counterfactual, we backtrack further, supposing that not only would it have been an 80s party, but also that Smith would have (contrary to how he actually is) been someone who liked 80s parties. Backtracking sometimes begets backtracking!

4.2 Default non-backtracking interpretations

Given the availability of these two interpretations, why are non-backtracking interpretations the default? With my semantics in hand, we can now reframe this question as follows: why is there a default preference to interpret a counterfactual as having a later admissible counterfactual time? In this section, I will offer an answer that makes crucial use of the structure of historical modality. In broad strokes, my explanation goes as follows. First, speaker intentions constrain the semantic values of context sensitive expressions in context, in particular past tense morphemes. Second, pragmatic assumptions about speaker intentions and relevant stereotypical background information often result in default interpretations of such context sensitive expressions.[36] By *default* here I mean that these interpretations are innocent until proven guilty—these default interpretations are overridable only if there is sufficient information in the context to override the background assumptions that generate them. In the case of counterfactuals, the generalization about speakers is that they intend to speak truly; the general background stereotype about counterfactuals is that, since historical possibilities decrease over time and since counterfactuals are universal quantifiers over subsets of such possibilities, interpreting a counterfactual as having a later admissible counterfactual time will generally give it a better chance of being true. Together, these assumptions yield a default preference to interpret counterfactuals as having later admissible counterfactual times and hence as non-backtracking.

The first component of my argument for default later admissible counterfactual times is the following hypothesis:[37]

[36] See Sperber & Wilson 1986, Wilson & Sperber 2012, Levinson 2000, Carston 2002 for related proposals about how pragmatic assumptions may influence what is asserted by an utterance of some sentence.

[37] The assumption that speaker intentions are the only factor relevant to determining the value of context-sensitive expressions is quite controversial (see for instance Lewis 1979b, Richard 2004, Glanzberg 2007,

INTENTION: Speaker intentions determine the value for non-automatic context sensitive expressions in the context.

By *non-automatic* context sensitive expressions I mean expressions like demonstratives, as well as gradable adjectives, epistemic modals, and conditionals (among many others). These are expressions whose semantic value depends on context in a non-obvious way, in contrast to 'automatic' indexicals like 'I' and 'tomorrow' for instance (cf. Perry 1997). Following Kaplan 1989, it is widely agreed that the value for these non-automatic context sensitive expressions in context depends at least in part on the intentions of the speaker.

I put forward INTENTION as a reasonable first pass theory (modulo the complications noted in footnote 37) about how we interpret context sensitive expressions. For instance, if we setting up a party and you say:

(14) All the beer is in the fridge.

the most natural interpretation of what you said is that all the beer *that we bought for the party* is in the fridge. This observation is nicely explained by INTENTION. In this case, I will assume that you intend to say something relevant about our current topic of conversation ('are we ready to party?'), and thus infer that you intend your claim to be about the beer we bought for the party (and not, for instance, the beer that we did not buy, which is still on the shelves in the store).

The second component of my argument is that there are general assumptions about (a) speaker intentions and (b) relevant background information which (i) result in default interpretations for certain context sensitive expressions that are (ii) overridable only if there is sufficient information in the context to override the background assumptions that generate them. I intend this proposal to be a rough first-pass theory of how overridable default interpretations of context-sensitive expressions arise.[38]

Regarding (a), here are some examples of assumptions about speakers' intentions:

TRUTH: Generally, speakers intend to speak truthfully.

2015, Dowell 2012, 2013, King 2013b,a, 2014). For the sake of space and to keep the metasemantics as simple as possible, I will not motivate this assumption but merely flag here that the final theory will likely be more complex than the one I discuss below. Given that the metasemantics of context sensitive expressions remains in theoretical infancy, I think this a reasonable move to make at this point.

[38]For instance, I will not defend this theory over competing relevance theories (Sperber & Wilson 1986, Wilson & Sperber 2012, e.g.), which would do similar work by appealing entirely to considerations of relevance. These alternative theories may be adapted to do the kind of explanatory work I aim for regarding the default interpretations of counterfactuals.

INFORMATIVITY: Generally, speakers intend to be maximally informative in what they say.

RELEVANCE: Generally, speakers intend for what they say to be relevant to the topic of the conversation.

These generalizations will hold for most speakers in most contexts, for the usual Gricean reasons. As such, for any particular context, unless we have specific reason to think otherwise, we tend to assume that the speaker intends to speak truthfully, informatively, and relevantly.

Regarding (b) and (i), here are some examples of the kind of relevant background information which seems to lead to default interpretations of past tense. Consider the following contrast (cf. Taylor 2001, Wilson & Sperber 2012):

(15) a. I have not had breakfast.
b. I have not been to Paris.

Intuitively, by default, an utterance of (15-a) will be most naturally interpreted to mean that the speaker has not had breakfast *today*, whereas for (15-b) that the speaker has *never* been to Paris. These default interpretations are naturally explained by INTENTION, the general assumptions about speaker intentions (TRUTH and INFORMATIVITY), and certain default stereotypes.

Take (15-a). The claim that no breakfast-event occurred throughout some strictly larger interval is more informative than the claim that no breakfast-event occurred throughout the strictly smaller interval (since the former entails the latter). But the claim that no breakfast-event occurred throughout the strictly larger interval is more likely to be false than the claim that no breakfast-event occurred throughout the strictly smaller interval (since it is a stronger claim). So, considerations of informativity weigh in favor of interpreting the claim as about a larger interval, and considerations of truthfulness weigh in favor of interpretation the claim as about a smaller interval. Given our general assumption that speaker intends to be both truthful and informative, what breaks the tie? Most naturally, we also assume the general stereotype that people usually eat breakfast once per day (and usually in the morning). Given this stereotype, we assume that it is likely false that the speaker has not had breakfast in the past 48 hours (and so on for any larger past interval)—thus, we rule out that the claim is about a past interval extending earlier than today. As such, the interpretation of (15-a) that best weighs considerations of likely truth

and informativity is that the speaker has not had breakfast today. Hence, since we expect that the speaker to be truthful and informative, our default interpretation of her utterance of (15-a) is that she has not had breakfast today.

Now consider (15-b). As before, the claim that no Paris-event occurred throughout some strictly larger interval is more informative than the claim that no Paris-event occurred throughout the strictly smaller interval, but the claim that no Paris-event occurred throughout a strictly larger interval is more likely to be false than the claim that no Paris-event occurred throughout the strictly smaller interval (since it is a stronger claim). In this case, there is no general stereotype about when people (non-Parisians) usually go to Paris. As such, there is no general reason to think that it is any more unlikely that a particular person has never been to Paris than merely that she has not been to Paris today, or this month, or this year. Thus, since to say that one has never been to Paris is more informative than that one has not been to Paris this month (and so on), and since we expect that the speaker intends to be truthful and informative, our default interpretation of her utterance of (15-b) is that she has never been to Paris.[39]

Of course, these interpretations are mere defaults, and may be overridden. This brings us to (ii): such default interpretations are overridable, but only if there is sufficient information in the context to override the background assumptions that generate them. Let us consider two scenarios:[40]

`Scenario 1`. It is Saturday at 11:15am. You and your roommates Sue and Ben

[39]Notice that if we remove the negations, the default interpretations remain:

(i) a. I have had breakfast. (Today)
b. I have been to Paris. (At least once in my life)

We predict this as well. Take (i-a). The claim that some breakfast-event occurred throughout some strictly smaller interval is more informative than the claim that some breakfast-event occurred throughout the strictly larger interval (again, since the former entails the latter). But the claim that some breakfast-event occurred throughout the strictly smaller interval is more likely to be false than the claim that some breakfast-event occurred throughout the strictly larger interval. So, considerations of informativity weigh in favor of interpreting the claim as about a smaller interval, and considerations of truthfulness weigh in favor of interpretation the claim as about a larger interval. As before, the general stereotype that people usually eat breakfast once per day (and usually in the morning) helps break the tie, settling on an interpretation in which (by default), an utterance of (i-a) says that the speaker has had breakfast this morning. I will not also review how we predict a similar default interpretation of (i-b) but I think the reader has a sense of how that should go.

[40]There are many other scenarios to consider. For instance, if you know the stereotype that people eat breakfast once per day does not apply to Ben because he has a policy of only eating breakfast once per week, then we predict (correctly) that when he utters 'I have had breakfast' in such a context, he says that he has had breakfast this week.

are deciding whether to eat at Breakfast Express or Lunch @ Sal's. You and Ben stayed up all night talking from midnight to now, and neither of you ate anything during that time. So, you know that Ben has not had breakfast this morning, and you know that Ben knows this too. Sue then says, 'Ben has had breakfast.'

`Scenario 2.` It is Saturday at 11:15am. You and your roommates Sue and Ben are deciding whether to eat at Breakfast Express or Lunch @ Sal's. You and Ben stayed up all night talking from midnight to now, and neither of you ate anything during that time. So, you know that Ben has not had breakfast this morning, and you know that Ben knows this too. Ben then says, 'I have had breakfast.'

In `Scenario 1`, it seems to me that Sue says that Ben has had breakfast today, even though this is false. In `Scenario 2` it is unclear to me what Ben says: either he is saying that he has had breakfast today and is lying, or maybe he is saying that he has had breakfast this week, and is perhaps communicating that he does not want to eat breakfast more than once per week. My response to Ben in this scenario would be, 'Wait, you did not eat anything this morning. What do you mean you have had breakfast?'

Notice that in `Scenario 1`, although you have good reason to think that 'Ben has had breakfast' is false on its default interpretation, it has that interpretation as uttered by Sue nonetheless. Thus, merely thinking that a sentence is false on its default interpretation is not sufficient to override the default; after all, the speaker may simply be mistaken. This is predicted by our proto-theory above: thinking that the sentence is false on its default interpretation is consistent with the assumption that the speaker aims to speak truthfully, and is not evidence that the subject does not conform to the relevant background stereotype (in this case, that Ben eats breakfast once per day), so we predict that the sentence ought to continue to have its default interpretation.

The crucial difference in `Scenario 2` is that you know that Ben knows that he has not had breakfast today. Knowing that the speaker knows that the default interpretation of his sentence is false undermines interpreting that sentence by its default, if we continue to assume that the speaker aims to speak truthfully. If we have good reason to continue to assume that the speaker aims to be truthful, then we look for an alternative non-default interpretation. However, since there is no obvious alternative interpretation, this leads us to wonder what the speaker meant. Alternatively, we might interpret the sentence on its

default interpretation and thus infer that the speaker is lying.[41] Since the information in the context does not settle which interpretive option is correct, we predict that this will lead to ambivalence in how Ben is interpreted in `Scenario 2`.

Let us now turn to apply this proto-theory of default interpretations to counterfactuals. We begin with the observation that historical possibilities decrease over time (recall Figure 1):

(16) For any world w and times t, t' such that $t < t'$: the historical possibilities at w, t' are a subset of the historical possibilities at w, t.

Furthermore, since for any A, w, t: $D_c(\mathsf{A}, w, t)$ is a subset of the historical possibilities of w, t, this feature of historical possibilities supports the general heuristic that counterfactual domains decrease over time:[42]

HEURISTIC: Generally, if $t < t'$, then $D_c(\mathsf{A}, w, t') \subseteq D_c(\mathsf{A}, w, t)$

Consider an utterance of some counterfactual $A \boxdot\!\!\rightarrow C$, and suppose that you qua hearer do not know, for any t, whether $D_c(\mathsf{A}, w, t) \subseteq \mathsf{C}$. Then, given HEURISTIC, for any two admissible counterfactual times t and t' such that $t < t'$, you should think it more likely that all worlds in $D_c(\mathsf{A}, w, t')$ are C-worlds than that all worlds in $D_c(\mathsf{A}, w, t)$ are C-worlds. Hence, in such circumstances, you should think that an utterance of $A \boxdot\!\!\rightarrow C$ has a better chance of being true if it has a later admissible counterfactual time than if it has an earlier one.[43] Given a historical-modality theory of counterfactuals like mine, this is the general stereotype relevant for interpreting any counterfactual.

As such, we expect that TRUTH + HEURISTIC will generate pressure to interpret counterfactuals as having later admissible times. But what about INFORMATIVITY and RELEVANCE? RELEVANCE constrains the interpretation of certain counterfactuals to share domains with others in the context—this is something we will come back to in §4.3. As

[41] Notice that even though you may suspend the assumption that Ben aims to speak truthfully in this case, this does not undermine his sentence having its default interpretation. This is because other members of the conversation (in this case, Sue) will continue to make that assumption and thus opt for the default interpretation. This is an instance of the general fact that lying exploits the assumption of truthfulness imparted to speakers; notice in particular that the speaker lying must assume that his interlocutors believe him to be telling the truth, else his intention to deceive would be foreseeably frustrated.

[42] Indeed, the only case in which this heuristic fails is if just the right true propositions are salient in the context in which the counterfactual is evaluated. Furthermore, knowing which ones must be salient is extremely unlikely, which is why it will almost invariably be assumed that HEURISTIC holds universally.

[43] This is so even if we assume a selection function semantics (a la Stalnaker), for the fewer possibilities in the counterfactual's domain, in principle the fewer opportunities to yield indeterminacy rather than truth.

such, RELEVANCE will not play any role in predicting the default interpretation of counterfactuals. Regarding INFORMATIVITY, my semantics predicts that there will be no general relationship between the choice of counterfactual time and informativity. This is because differences in the causal sufficiencies and salient post-counterfactual-time facts will yield non-entailments in both directions of time (I demonstrate this in the Appendix). Since we do not generally evaluate counterfactuals with full knowledge of the relevant causal sufficiencies and salient post-t facts, it will generally be the case that for any times $t < t'$, the interpretation of $A \boxrightarrow C$ with counterfactual time t will not strictly entail (nor be strictly entailed by) the interpretation of $A \boxrightarrow C$ with counterfactual time t'. Therefore, INFORMATIVITY will generally not constrain the interpretation of counterfactuals.

Hence, INTENTIONS, together with HEURISTIC and our general expectations that the speaker intends to speak truthfully (TRUTH), be informative (INFORMATIVITY), and relevant (RELEVANCE) generates the following default interpretation for counterfactuals:

> **Default**: Generally and by default, the counterfactual time for $A \boxrightarrow C$ is some admissible time t such that for any later admissible time t', $D(\mathsf{A}, w, t) \subseteq \mathsf{C}$ iff $D(\mathsf{A}, w, t') \subseteq \mathsf{C}$.

Notice that, once you reach some admissible time t such that the counterfactual's truth does not depend on whether its counterfactual time is t or any later admissible counterfactual time, then INTENTIONS, HEURISTIC, TRUTH, INFORMATIVITY, and RELEVANCE provide no additional default constraint on which time in that range is *the* counterfactual's evaluation time.[44]

[44] I pause to point out two important upshots of **Default**. The first is that **Default** is compatible with time being dense and hence with the possibility that there are later and later admissible counterfactual times without end. The second is that it predicts that by default many counterfactuals will be indeterminate, since once we reach a late admissible time such that the counterfactual's truth value is invariant as evaluated with respect to it or any later admissible time, **Default** provides no further constraint on counterfactual time. This yields the prediction that when we evaluate a non-backtracking counterfactual, there is a certain degree of leniency in how much of the past we hold fixed. Take (3) for instance:

(3) If Nixon had pressed the button, there would have been a nuclear holocaust.

By default, we hold fixed much of the past in evaluating this counterfactual: in particular, we hold fixed that the button is wired to the launching mechanism and so on. However, there remains quite a bit of indeterminacy in just how much of the past we hold fixed in evaluating (3). For instance, it seems indeterminate whether Nixon's button-pressing event itself a miracle (thus holding fixed all of the past up until that moment) or whether some small earlier miracle leads lawfully to Nixon's pressing the button (thus holding fixed less of the past). Such indeterminacy is correctly predicted by **Default** (cf. Lewis 1979a: 463, Lewis 1981: 118). We will discuss backward counterfactuals whose consequents state miraculous events in §4.2.2.

Crucially, since later admissible times give rise to non-backtracking interpretations, **Default** predicts that non-backtracking interpretations are the default. Thus, it predicts (B): the default interpretation of a forward counterfactual is non-backtracking. Tying things back to (1), we predict that it will be interpreted (by default) as having t_2 as its counterfactual time (from our diagram Figure 4)—that is, a time at which it is historically settled that Smith was rational, wanted to live, and believed jumping would kill him, and at which it is historically settled that there was no net underneath him right before his jump. Hence, on its default interpretation, (1) has a later admissible counterfactual time, and thus is true.

I pause to anticipate the following challenge. It may seem that my theory also predicts that an utterance of (2) without any preamble (for instance, prior to Beth's speech) should also be interpreted so that it is true, for exactly the same reason as (1).

(2) If Smith had jumped, he would have lived.

Someone may reason like this: the speaker aims to speak truly (given TRUTH), so she cannot have intended that (2) have its default counterfactual time (since thus interpreted, it would be false). Therefore, she must have intended some other, earlier counterfactual time, and spoken truly. This is a terrible prediction: without the extra context provided by Beth's speech, (2) seems to have only the false non-backtracking interpretation; we do not go looking for some backtracking interpretation on which it is true.

Contrary to the thought expressed in the previous paragraph, my theory does not predict that we will automatically judge a context-initial utterance of (2) true. Recall observation (ii): default interpretations are overridable, but only if there is sufficient information in the context to override the background assumptions that generate them. In particular, knowing that S is false on its default interpretation is not sufficient to think that it has a non-default interpretation. We saw this above with `Scenario 1`: merely knowing that Ben had not had breakfast today is not sufficient to lead us to think that Sue's utterance of 'Ben has had breakfast' has a non-default interpretation. Likewise, merely knowing (or thinking it likely) that (2) is false on its default interpretation is not sufficient to lead us to think that it has a true non-default interpretation. In both cases, we assume that the speaker, though aiming to speak truthfully, fails to do so.

Now, if we also have reason to think that Beth also knows that (2) is false on its default interpretation, then my intuition changes. It is a bit tricky to set up a context in which

this is clear, but here is one:

(17) Beth: I also agree that there was in fact nothing underneath Smith to break his fall. Nonetheless, if he had jumped, he would have lived.

As in `Scenario 2`, I find such an utterance by Beth puzzling, and in this case I would try to find a non-default interpretation of her utterance of (2). Since none are obvious, I might ask for clarification about what Beth meant, prompting something like Beth's backtracking speech. The crucial point here is that my theory predicts that we will not interpret (2) as having a non-default (backtracking) interpretation simply because we think it is false on its default (non-backtracking) interpretation. It does predict that if we think the speaker thinks (2) is false on its default interpretation, and if there is no salient intended non-default interpretation, our reaction to the utterance will be one of unclarity rather than a judgment of falsity. Both predictions seem to be accurate.

Before we move on to predicting (C), I want to address two remaining concerns with my explanation of (B).

4.2.1 Might-counterfactuals?

My explanation of (B) is that, since counterfactuals are universal quantifiers (or choice functions over a particular domain), smaller domains have a better chance of making the counterfactual true; thus, since later times generally give rise to smaller domains (by HEURISTIC) and speakers generally intend to speak truthfully (TRUTH), non-backtracking interpretations are the default. However, *might*-counterfactuals seem to be most naturally interpreted as existential quantifiers over the same domain, as follows:

(18) $A \diamond\!\!\rightarrow C$ is true at c, w, t iff some world in $D_c(\mathsf{A}, w, t)$ is a C-world.

Thus, by the above reasoning, we expect the default interpretation of *might*-counterfactuals to be backtracking, since we expect such interpretations to give them a better chance at being true. But, so the problem goes, *might*-counterfactuals do not have default backtracking interpretations.[45]

In response, I accept that my view has this consequence for *might*-counterfactuals whose truth conditions are given by (18); however, I want to mitigate the apparent badness of the result by holding that *might*-counterfactuals are actually ambiguous between that reading

[45] Thanks to an anonymous reviewer for raising this challenge.

and one on which they rather express the epistemic possibility of the corresponding *would*-counterfactual, $\Diamond(A \boxright C)$ (cf. Stalnaker 1980, DeRose 1999). Many of the arguments for this view take a stand on issues that I prefer to remain neutral on (for instance the Limit and Uniqueness Assumptions, and Conditional Excluded Middle). However, at least one of Stalnaker's arguments for the possibility of this reading carries over (see pp. 99-100). If (18) gives the truth-conditions for every *might*-counterfactual and we adopt a universal quantifier semantics for *would*-counterfactuals, then we should expect utterances of sentences like *If A, it might be that not-B, although I believe that if A then it would be that B* to be infelicitous. The reason is that, given these assumptions, the first counterfactual should be equivalent to the negation of the second counterfactual, and thus utterances of such sentences should be as infelicitous as utterances of sentences like, *Not-A, although I believe that A*. However, utterances of sentences of the first kind are not always infelicitous:

(19) If Smith had jumped, he might have lived, although I believe that if Smith had jumped, he would have died.

If we instead analyze the *might*-counterfactual here as involving epistemic *might* on top of the corresponding *would*-counterfactual, then we predict that (19) should be just as felicitous as an utterance of a sentence like, *It might be that not-A, although I believe that A*. For instance, (19) should be as felicitous as (the switch to present tense is to force the epistemic interpretation of *might*):

(20) Smith might be upstairs, although I believe he is outside.

Of course, the kind of reading I am positing for some *might*-counterfactuals raises tricky issues in the compositional semantics of counterfactuals (see for instance Swanson 2011) which I want to set aside for now. However, even granting me this point, one might rephrase the worry as follows. Since I grant that there is an interpretation of *might*-counterfactuals on which their truth conditions are given by (18), we can rephrase the challenge by targeting exactly such a counterfactual. Consider the following example of Lewis, which seems to clearly violate the epistemic *might* interpretation. Smith in fact has only dimes in his pocket, but he does not know this and he says:

(21) If I had looked in my pocket just now, I might have found a penny.

Since for all Smith knew, *if Smith had looked in my pocket, he would have found a penny*

is true, it seems the *might*-counterfactual ought to be true, on the wide-scope epistemic *might* interpretation. But intuitively, (21) is false, merely because there was no penny in Smith's pocket. Let us accept this judgment for the time being (though see DeRose 1994 for dissent). This may be seen as a serious problem for my view, because according to my view, (21) ought by default to have a backtracking interpretation, but it seems to have a non-backtracking interpretation by default.

To this challenge, I respond that it is not so obvious that (21) does *not* have a default backtracking interpretation, even one on which its counterfactual time is before it is historically settled that there are any pennies in Smith's pocket. Remember that on my theory, the counterfactual's domain will entail any post-t salient facts not 'disrupted' by the supposition of the antecedent (via HINDSIGHT). Since looking in my pocket will not disrupt the fact that there is no penny there (at least we are given no reason to think it would), and since this fact is very clearly salient in the context, my theory predicts that it will be entailed by (21)'s domain, even on such a backtracking interpretation, and hence it will come out false even on that interpretation. Therefore, (21) *may* in fact have a backtracking reading by default—our intuitions about its falsity do not tell one way or the other.

Of course, to defend against one alleged counterexample is not to provide any positive reason to think my theory's prediction here is correct. Thus, I close this section with a short, and admittedly inconclusive, piece of data in favor of default backtracking interpretations of *might*-counterfactuals. Consider the scenario described at the outset of the paper, with Smith just seconds previously poised to jump. Beth says,

(22) If Smith had jumped, he might have lived.

Just prior to hearing Beth's utterance, I am prepared to accept (1)—if Smith had jumped, he would have died. However, my reaction to Beth's utterance is *not* that it is obviously false. Rather, I am quizzical. I want to know why she thinks that. If Beth says something like, 'Well, there are all sorts of reasons he might have jumped; maybe he would have been wearing a parachute or maybe he would have had a net underneath him,' this makes it clear that she intended (22) on its non-epistemic interpretation (with truth conditions given by (18)), and it seems most clear that it has a backtracking interpretation. If Beth instead says something like, 'Well, there are all kinds of things that might have happened, so I do not know what would have happened,' this makes it clear that she intended it on its epistemic interpretation, with the embedded *would*-counterfactual having its default

non-backtracking interpretation.

I conclude that there are viable responses available in defense of my theory regarding *might*-counterfactuals.

4.2.2 Latest default admissible counterfactual time?

The final worry I will consider about my theory's prediction of default non-backtracking interpretations has to do with its prediction that the default counterfactual time is its latest admissible counterfactual time (**Default**). The worry is that **Default** will force late counterfactual times and hence lead my theory to false predictions about certain backward counterfactuals. The following example from Bennett 2003 (pp. 209-10) illustrates the problem:

> `Dam`. A dam suddenly bursts, quickly submerging a low-lying road and killing the people stuck in their cars in its path.

In `Dam`, the following counterfactual seems true:

(23) If there had been no cars on the road just then, no lives would have been lost.

However, in accepting this, we are clearly not thereby committed to the cars that were actually on the road suddenly and miraculously ceasing to be there, just before whatever time 'then' picks out. If that were so, accepting (23) would incorrectly commit us to a counterfactual like the following being true:

(24) If there had been no cars on the road just then, they would have suddenly ceased to be there.

(I pause to flag that Lewis 1979a's theory is well-positioned to predict the truth of (23) and falsity of (24) in `Dam`. According to Lewis, it would take a larger miracle (or set of miracles) to make all the cars there to vanish just as the dam break than it would to make each of the drivers of those cars to decide (at various earlier times) not to drive them onto that road. As such, Lewis predicts that (23) may be true, though (24) is false.)

How, then, can my theory, which is committed to **Default**, predict that counterfactuals like (24) are not true? I will argue here that, given **Contingent consequent**, the only admissible counterfactual times for (24) are ones on which it is false. Begin by noticing that the event described by the consequent of (24) is an event comprising at least two

points of time: a time at which the cars are located on the road and a time immediately after at which the cars are no longer on the road. Let E be the proposition that an event of the cars on the road at t_1 suddenly ceasing at t_2 to be located on the road occurred $(\mathsf{C}_{t_1} \wedge \overline{\mathsf{C}}_{t_2})$. Finally, let $\overline{\mathsf{C}}_{t_2}$ be the proposition expressed by (24)'s antecedent—that there are no cars on the road at t_2.

Given **Contingent consequent**, my theory predicts that (24) must be false. Here is why. By **Contingent consequent**, an admissible counterfactual time for (24) must be some time t such that $H(w,t) \cap \overline{\mathsf{C}}_{t_2} \cap \mathsf{E} \neq \emptyset$ and $H(w,t) \cap \overline{\mathsf{C}}_{t_2} \cap \overline{\mathsf{E}} \neq \emptyset$. However, at any such time, (24) will be false! Quick proof:

The only admissible counterfactual times for (24) are before t_1.

- Suppose (24)'s counterfactual time is t_1, or any time after t_1. Let t^+ be an arbitrary such time. Then $H(w,t^+) \cap \overline{\mathsf{C}}_{t_2} \models \mathsf{C}_{t_1}$, since C_{t_1} is true at w and about times no later than t^+ (by HISTORICAL). Therefore, $H(w,t^+) \cap \overline{\mathsf{C}}_{t_2} \models \mathsf{E}$ (remember, $\mathsf{E} = \mathsf{C}_{t_1} \wedge \overline{\mathsf{C}}_{t_2}$). Hence $H(w,t^+) \cap \overline{\mathsf{C}}_{t_2} \cap \overline{\mathsf{E}} = \emptyset$. Therefore, t^+ is not an admissible counterfactual time for (24).
- Suppose instead that counterfactual time is before t_1. Let t^- be an arbitrary such time. Then $H(w,t^-) \cap \overline{\mathsf{C}}_{t_2} \not\models \mathsf{C}_{t_1}$. Hence, $H(w,t_1) \cap \overline{\mathsf{C}}_{t_2} \cap \overline{\mathsf{E}} \neq \emptyset$. But equally, since $H(w,t^-) \cap \overline{\mathsf{C}}_{t_2} \not\models \overline{\mathsf{C}}_{t_1}$, $H(w,t_1) \cap \overline{\mathsf{C}}_{t_2} \cap \mathsf{E} \neq \emptyset$. Thus, t^- is an admissible counterfactual time for (24).

But now for any time $t < t_1$, if t is the counterfactual time for (24), it is false.

- To see why, first notice that $D_c(\overline{\mathsf{C}}_{t_2}, w, t) \models (\mathsf{C}_{t_1} \supset \mathsf{C}_{t_2})$; this is because $\mathsf{C}_{t_1} \supset \mathsf{C}_{t_2}$ is a causal sufficiency of w whose antecedent is about times extending beyond t^-. Therefore, since $D_c(\overline{\mathsf{C}}_{t_2}, w, t) \models \overline{\mathsf{C}}_{t_2}$, it follows via modus tollens that $D_c(\overline{\mathsf{C}}_{t_2}, w, t) \models \overline{\mathsf{C}}_{t_1}$. But then $D_c(\overline{\mathsf{C}}_{t_2}, w, t) \models \overline{\mathsf{E}}$.

Now, my way of predicting (24) may seem to rely on a trick involving 'ceasing to be' denoting an event which has a temporal extension. In fact, it does not. Rather, it relies on the fact that **Contingent consequent** drives counterfactual time to be earlier than consequent time, and the fact that CAUSAL demands that counterfactual domains entail causal sufficiencies with antecedents about times extending beyond counterfactual time. Let us thus consider a related problematic example in which the counterfactual's consequent is about a momentary event or state, such as:

(25) If there had been no cars on the road at t_2, there would have been cars on the road just before then at t_1.

This seems false, just as (24). So how do predict this? Again, by **Contingent consequent**, the counterfactual time for (25) must be before the time its consequent is about, hence before t_1.[46] Let us suppose (for the hardest case) it is the time just prior, t_0. By HISTORICAL, $D_c(\overline{\mathsf{C}}_{t_2}, w, t_0)$ entails that the cars are on the road at t_0, since they are on the road at t_0 at w. As before, by CAUSAL, $D_c(\overline{\mathsf{C}}_{t_2}, w, t_0) \models \mathsf{C}_{t_1} \supset \mathsf{C}_{t_2}$. But of course $D_c(\overline{\mathsf{C}}_{t_2}, w, t_0) \models \overline{\mathsf{C}}_{t_2}$. Therefore, by modus tollens, $D_c(\overline{\mathsf{C}}_{t_2}, w, t_0) \models \overline{\mathsf{C}}_{t_1}$. Hence, we predict that (25) is false.

I conclude that **Default** is not threatened by examples like these—my theory predicts correctly that such counterfactuals are false. However, I pause to briefly articulate one upshot of this result. First, notice that (24)'s consequent is the negation of some causal sufficiency of w, $\mathsf{C}_{t_1} \supset \mathsf{C}_{t_2}$. Therefore, just as my theory predicts that conditionals with sufficiency-violating consequents must be false, it predicts that conditionals whose consequents state true causal sufficiencies must be true. But now suppose that determinism is true and that causal sufficiencies are instances of natural laws, and consider an arbitrary counterfactual $A \boxdot\!\!\rightarrow C$ whose consequent is not about any causal sufficiency (for instance, (23) above). Then, my theory predicts that this counterfactual's domain $D_c(\mathsf{A}, w, t)$ will contain worlds which violate some of w's causal sufficiencies. Nonetheless, when we try to state what such violations would have been using a backward counterfactual $A \boxdot\!\!\rightarrow M$, by **Contingent consequent**, the domain of this counterfactual $D_c(\mathsf{A}, w, t')$ will comprise only worlds which do not violate that sufficiency. Sufficiency-violations (or miracles, in Lewis's terminology) are thus predicted by my theory to be elusive.[47] We know counterfactual domains must comprise worlds where miracles occur by appreciating how the semantics works, but we cannot use counterfactual language to state that they are there.

[46] As above, suppose (25)'s counterfactual time is any time t_1 or later. Then $H(w, t_1) \cap \overline{\mathsf{C}}_{t_2} \models \mathsf{C}_{t_1}$, since C_{t_1} is true at w and about times no later than t_1 (by HISTORICAL). Therefore, $H(w, t_1) \cap \overline{\mathsf{C}}_{t_2} \cap \overline{\mathsf{C}}_{t_1} = \emptyset$. Therefore, any time t_1 or later is not an admissible counterfactual time for (25).

[47] This is related to Lange 2000's idea that miracles are 'off stage', though Lange is primarily interested in preserving the intuition that laws survive counterfactual suppositions. I can see various ways of connecting these two ideas, but I will not pursue this project at this time.

4.3 How to backtrack

As I mentioned above, **Default** can be overridden in certain contexts, if there is sufficient information in the context to determine which non-default interpretation the speaker intends. In this section, I will sketch an account of one way **Default** may be overridden, thus giving rise to backtracking interpretations. Since I aim to explain (C), I will focus on how an assertion of a backward counterfactual is sufficient to override **Default** and bias a backtracking interpretation of a related forward counterfactual.

C. Asserting a backward counterfactual will often make salient a backtracking interpretation of a forward counterfactual sharing the same antecedent with, and uttered after, that backward counterfactual.

Recall from §4.1 how my semantics predicts that the counterfactual (2) will receive a true backtracking interpretation when uttered in the context of Beth's speech and her prior assertion of (9):

> 'Smith was rational, had no wish to die, could see below him, and knew that jumping without a net would kill him. Therefore,
>
> (9) Had Smith jumped, there would have already been a net below him to catch him safely.

So,

> (2) If Smith had jumped, he would have lived.'

In particular, recall that we predict that (9) is true if its counterfactual time is some late admissible time t at which (by **Contingent consequent**) its consequent is not J-historically settled and its domain entails the salient post-t propositions that Smith was rational, wanted to live, and knew whether there was a net beneath him. Furthermore, we predict that (2) has a true backtracking interpretation if it has the same (in this case, earlier than default) counterfactual time and its domain also entails the same salient post-t propositions as (9). However, we have not yet seen why, in the context of Beth's speech, it should be the case that:

α. (9) is interpreted with respect to a set of salient propositions that includes that Smith was rational, wanted to live, and knew whether there was a net beneath him, and

β. (2) has the same counterfactual time as (9), and is interpreted with respect to a set of salient propositions that also includes that Smith was rational, wanted to live, and knew whether there was a net beneath him.

In the rest of this section, I will motivate (α) and (β).

(α): What are the salient propositions?

I do not have a fully general answer to the question of what post-t propositions will be contextually salient in a particular context, and I will not offer an account of what contextual salience *is*. However, this should not dissuade us from appealing to a notion of salience in stating our semantics (as I have done with HINDSIGHT). The reason is that we can say enough about the intended notion of salience so that the resulting theory makes testable predictions. For instance, it seems obvious that asserting a proposition is sufficient to raise it to salience (at least for a short while). Notice that, with just this obvious thought in hand, our theory makes several interesting (and by my intuitions, correct) predictions. For instance, combining this observation with HINDSIGHT yields the prediction that (9) will seem true given the preamble in **A** but not given the preamble in **A′** (notice that the preamble in **A** is Beth's speech from earlier):

A. Smith was rational, had no wish to die, could see below him, and knew that jumping without a net would kill him. Therefore, **had Smith jumped, there would have already been a net below him to catch him safely.**

A′. There was nothing below Smith to break his fall in the event of a jump. Therefore, ??**had Smith jumped, there would have already been a net below him to catch him safely.**

While (9) seems natural and acceptable in context **A**, it is decidedly less natural and acceptable in the context **A′**. Asserting that there was nothing below Smith to break his fall makes it hard to hear a subsequent assertion of (9) as true; furthermore, an assertion of this string of sentences is odd (perhaps because it is unclear why someone would deliberately set themselves up to say something obviously false).[48] The contrast between **A** and **A′** is evidence that whether we find a backward counterfactual acceptable (or felicitous)

[48]Indeed, notice that even just asserting that there was nothing below Smith in addition to the other assertions in context **A** results in an odd string:

(i) Smith was rational, had no wish to die, could see below him, and knew that jumping without a

sometimes depends on what is asserted before it in the context, and thus (assuming that asserting a proposition makes it salient) confirmation of HINDSIGHT. Furthermore, this evidence is not cherry-picked. As pointed out in the introduction, all the classic discussions of backtracking counterfactuals involved examples in which the backtracking interpretation was promoted by way of a contextual preamble, followed by the assertion of a relevant backward counterfactual (cf. Downing 1959, Jackson 1977, Lewis 1979a, Bennett 1984, 2003). Our observation about salience, plus HINDSIGHT, predicts the first half of this pattern (the contextual preamble before the backward counterfactual)—the second half is predicted by **Contingent consequent** (which ensures an earlier counterfactual time for the backward counterfactual) and (β). We turn next to explaining (β).

(β): Why are counterfactual times and salient propositions inherited?
To begin, recall observation (ii): default interpretations are overridable only if there is sufficient information in the context to determine which non-default interpretation the speaker intends. My proposal is that uttering a backward counterfactual and then a forward counterfactual with the same antecedent in close succession provides enough information in the context for your addressees to work out that you intend the forward counterfactual to share its counterfactual time with, and be evaluated with respect to, the same salient propositions as the backward counterfactual. The reason doing so provides the requisite information is that (i) it is generally assumed that the speaker intends her utterance to be relevant (RELEVANCE), and (ii) the most natural way to interpret back-to-back utterances of counterfactuals with the same antecedent as relevant is as sharing domains.

Here is an example of a similar phenomenon (that admits of a similar explanation) exhibited by nominal quantifiers. Recall the example from before, where we are setting up the party, and you say,

---

net would kill him. Furthermore, there was nothing below Smith to break his fall in the event of a jump. Therefore, ??**had Smith jumped, there would have already been a net below him to catch him safely.**

Interestingly, changing which propositions are asserted seems to bias us to favor alternative backward counterfactuals (something also predicted by my theory). For instance, I find the following string perfectly felicitous, and am inclined to accept the counterfactual as true as uttered in that context:

(ii) Smith was rational, could see there was no net beneath him, and knew that jumping without a net would kill him. Therefore, **had Smith jumped, he would have wanted to die.**

(14) All the beer is in the fridge.

As before, considerations of relevance naturally lead to interpreting you as meaning that all the beer that we bought for the party is in the fridge. Suppose now that Sue follows up, saying,

(26) Most of the beer is cold, but some is still warm.

Most naturally, Sue is talking about the same beer as you are—the beer that was bought for the party. RELEVANCE also predicts this observation. For what Sue says to be relevant, her claim should be about the same beer that John is talking about, since that beer is now the topic of conversation. Assuming RELEVANCE, we think that Sue intends for her utterance to be relevant and hence that she is talking about said beer.[49]

Now, on my theory, a counterfactual $A \boxrightarrow C$ is a claim about a domain of worlds $D_c(\mathsf{A}, w, t)$, just as John's utterance of (14) is about a domain of beer. So, just as a domain of beer can be a topic of conversation, so should a domain of counterfactual possibilities be a possible topic of conversation. Then, just as the relevance of Sue's utterance of (26) requires that it have the same domain as (14) (as uttered by John), the relevance of an utterance of (2) immediately following an utterance of (9) requires that their domains be the same. I now spell out the reasoning explicitly, using Beth's utterance as an example.

Initially, you say (1):

(1) If Smith had jumped, he would have died.

This has its default interpretation, which is non-backtracking and hence true. However, Beth counters by first asserting R, D, and K (hence raising them to salience) and then uttering the backward counterfactual (9):

(9) Had Smith jumped, there would have already been a net below him to catch him safely.

Since the default counterfactual time for (1) is t_2, and this is not an admissible counterfactual time for (9) (since t_2 is after it is settled that there was no net below Smith prior to his jumping), we cannot nontrivially interpret (9) as about the same time as (1). Instead, we interpret (9) via **Contingent consequent** as about an earlier counterfactual time t_0,

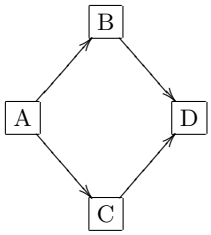
[49] I won't speculate about the best formal theory for modeling these phenomena. For a now-classic theory of how to represent the topics of a conversation, see Roberts 2012b,a.

and true (given the facts of the case). However, now Beth utters (2):

(2) If Smith had jumped, he would have lived.

Now notice that, by uttering (9), Beth makes clear a candidate non-default counterfactual time for (2). Assuming RELEVANCE, we naturally think that Beth's utterance of (2) is relevant, and hence about the same domain as (9) (since that's the only way in this context we can make sense of its being relevant). But since the salient propositions have not changed, the only way for the domain of these two counterfactuals to be the same is if they have same counterfactual time and are interpreted with respect to the same set of salient propositions. Thus, since t_0 is an admissible counterfactual time for (2), we interpret it as having that counterfactual time in this context, and also interpret it with respect to the same salient propositions as (9) ($\mathsf{R}, \mathsf{D}, \mathsf{K}$, all of which by HINDSIGHT will be entailed by $D_c(\mathsf{J}, w, t_0)$)—thus, we interpret (2) as true! Importantly, the utterance of (9) immediately prior to (2) is what allows Beth to make explicit the counterfactual time she intends for the latter, and hence why **Default** may be overridden in this case.[50]

[50] I remain neutral about other ways **Default** may be overridden to yield a backtracking interpretation of $A \boxrightarrow C$. However, my theory suggests that if there is enough information in the context to work out the speaker's intended non-default counterfactual time, and reason to think that the counterfactual is true so-interpreted, this ought to be enough to ensure the counterfactual has a backtracking interpretation (this is just a suggestion, because there may be other constraints on backtracking interpretations we have not explored). Hence, my theory is compatible with counterfactuals having backtracking interpretations even without a previous utterance of a backward counterfactual sharing the same antecedent. A casual reflection on simple examples like the Gerstenberg model suggests that this is correct.



Suppose that we both know the causal model. I then say to you:

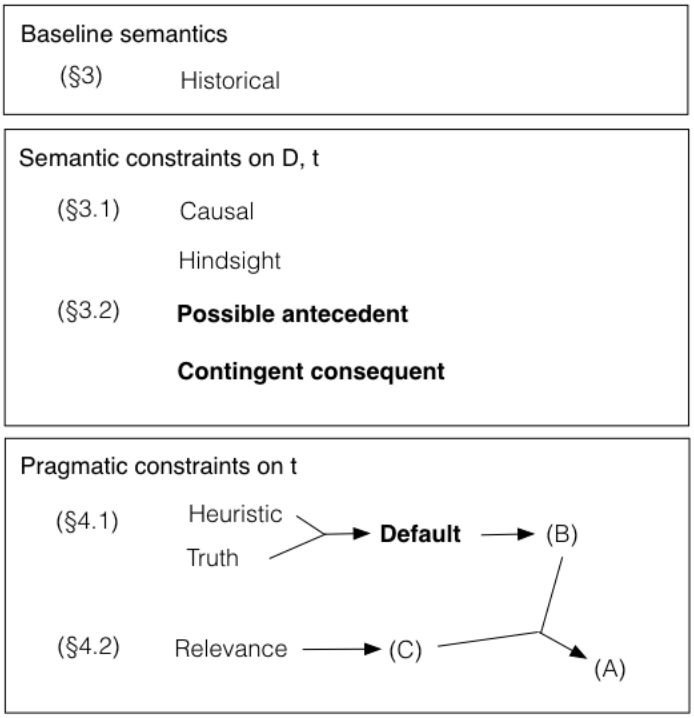
(i) If B had not happened, then D would not have happened either.

I submit that it is not too difficult to interpret (i) as backtracking and hence true. If I emphasize the lawful connections between A, B, C, and D, this also seems to help such an interpretation (as reported by several colleagues). My theory predicts this. We both know the causal model. Thus, we both know that (i) is obviously false on its default interpretation. Furthermore, it is easy to work out what my intended counterfactual time for (i) must be (some time before A). On that interpretation, we both know that (i) is true. Therefore, given TRUTH, my theory predicts that such a backtracking interpretation of (i) ought to be available and preferred.

That is my official line about how my theory predicts (C). We predict the possibility of backtracking counterfactuals because in certain contexts there will be enough information to work out the speaker's intended non-default counterfactual time. Furthermore, utterances of backward counterfactuals are a way of making explicit your intended non-default counterfactual time, hence (C).

Recap

Let us briefly recap the theory. I include the following graphic to show the building blocks of the theory and what is explaining what:



The two general features sufficient for accepting a counterfactual $A \boxdot\!\!\rightarrow C$ on its backtracking interpretation are:

- Enough information to work out the speaker's intended non-default counterfactual time t.
- Some reason to think that every world in $D_c(\mathsf{A}, w, t)$ is a C-world

Consider again Jackson's example from the outset of the paper. The default interpretation of (1) comes out true because on its default counterfactual time t, every world in $D_c(\mathsf{J}, w, t)$ is a world where there is no net beneath Smith when he lands and hence one in which his jump kills him. However, Beth's speech, plus her utterance of the backward counterfactual (9), induce two contextual shifts. The first is that the propositions R, D, and K become salient. The second is that Beth makes clear her intended non-default counterfactual time for (2) is some earlier time t^*, which she does by uttering it immediately after uttering (9). Thus, if Beth's assertion of (9) is accepted, it becomes accepted that every world in $D_c(\mathsf{J}, w, t^*)$ is ones in which there is a net beneath Smith at the time he lands, and hence one in which the jump does not kill him. Thus, once (9) is accepted, as long as RELEVANCE-considerations pressure us to interpret (2) as uttered by Beth just after as having the same domain as (9), it will also be accepted, so-interpreted.

5 Objections and replies

5.1 Special backward morphology?

Objection: *You have focused on backward counterfactuals, like (9), with normal counterfactual morphology. But some, such as (27), involve an extra 'have to' in their consequents. What is the difference, if any, between these backward counterfactuals, and can your theory predict this?*

(9) Had Smith jumped, there would have already been a net below him to catch him safely.

(27) Had Smith jumped, there would <u>have to have</u> already been a net below him to catch him safely.

Reply: I am not sure what the correct descriptive generalization about (9) and (27) is, let alone what best explains it. Nonetheless, there are good reasons to think that my theory is compatible with this data. Here is some reason to think that (27) is related to (9) in the way that strong necessity modal claims ('You have to wash your hands') are related to weak necessity modal claims ('You ought to wash your hands'):

> `Bet`. John is betting on horse races. The way the payout works is that those betting on the top three horses win a percentage on top of their bet. John in

fact bet on Slowmo, and Slowmo lost badly. Dasher, Dancer, and Prancer won. Right before the bet, John was torn between betting on Slowmo or betting on Dasher, but went with Slowmo at the last second.

In this scenario, I think that (28) is true but (29) is false:

(28) If John had won, he would have bet on Dasher.

(29) If John had won, he would have to have bet on Dasher.

In particular, I would describe the scenario as follows:

(30) If John had won, he would have bet on Dasher, but it's not the case that if he had won, he would have to have bet on Dasher.

This pattern seems analogous to me to the (deontic) 'You ought to wash your hands, although you do not have to.' That is, it is often possible to coherently assert a weak necessity modal claim conjoined with the negation of its strong necessity counterpart.

Explaining this fact about weak and strong necessity modals, and thus explaining what is going on with the extra 'have to' in (27), takes us far beyond the scope of this paper. But there is a promising proposal, due to von Fintel & Iatridou 2008, that can be easily adapted to my theory of counterfactuals. Their proposal is that 'ought' and 'have to' are both universal quantifiers, and that 'ought' quantifies over a subset of the possibilities that 'have to' does. Similarly, I could hold that the domain of ordinary backward counterfactuals is a subset of the domain of those with the extra 'have to'. Granted, more work remains for predicting such a result within my theory, but there is no reason to think that such an extension of my theory is impossible.[51]

5.2 Generality

Objection: *There are examples which suggest that the backtracking/non-backtracking split is a phenomenon which has nothing to do with time. Since your theory only captures tem-*

[51] Some people report that (27) sounds 'epistemic' in a way that (9) does not. That may be right, though my intuitions are not so clear. If it is right, this lends support to an alternative proposal to the one floated here, one which is equally compatible with my theory of counterfactuals. On this proposal, the extra 'have to' is an epistemic necessity modal which scopes above the counterfactual: $\Box(A \boxrightarrow C)$. (29) thus means something like: it must be the case that if John had won, he would have bet on Dasher. This is false because it is epistemically possible that if John had won, he would have bet on Dancer. For more on counterfactuals with underspecified antecedents, see §5.5.

poral examples, it does not account for the general phenomena.

Here is the example:[52] suppose Jesus walks by Miriam cooly. Then I say,

(31) If Jesus had pushed Miriam just then, he would have done something morally wrong.

You object, saying, 'Look, had Jesus pushed Miriam just then, the moral laws would have been different than they are (since Jesus is by definition morally perfect), so

(32) If Jesus had pushed Miriam just then, he would not have done anything morally wrong.'

(31) and (32) do not seem to differ in their counterfactual time, and even if they did, it is hard to see how that could be relevant, since what is morally wrong does not change across time. Yet, (31) is intuitively true, while (32) is intuitively true in the context of the objection (after the backward counterfactual has been asserted). Furthermore, the example seems analogous to (1)/(2) from the introduction. (32) seems true because when one 'backtracks' to make its antecedent is true, one has to change the moral laws such that Jesus pushing Miriam is not morally wrong.

Reply: Here are two responses. The first is to deny that the moral laws are contingent, and hence reject the intuition that both (31)/(32) are true in their respective contexts.[53] I take it that this response is reasonable, but does not quite cut to the heart of the matter. The second response is to accept the intuitions, but respond that although (31) and (32) do not *seem* to differ in counterfactual time, they in fact do. But then what counterfactual time might (32) have, in light of the assumption that what is morally wrong does not change across time? One possibility is to secure moral contingency by considering alternative branches from some initial moment of time. On this response, (32) has as its counterfactual time the initial moment of time (or at least some very early moment), at which it is historically contingent what the moral laws are. This picture is consistent with the moral laws being eternal in the sense that, once it is historically settled what the moral laws are, it is forever historically settled what the moral laws are. I submit that this way

[52] I owe this example to Tom Dougherty (personal communication).

[53] Though see Rosen 2014 for reasons to take seriously the idea that the fundamental moral laws are contingent.

of handling (31)/(32), although admittedly surprising, is not obviously false.

5.3 Disagreement

Objection: *According to the theory presented here, the content of a non-backtracking interpretation of* $A \boxrightarrow C$ *and a backtracking interpretation of* $A \boxrightarrow \neg C$ *do not contradict each other. Hence, the theory predicts no disagreement between you (asserting (1)) and Beth (asserting (2) after her preamble). Yet, intuitively the two of you* **do** *disagree. For instance, Beth can appropriately say 'no' in response to your utterance.*[54]

Reply: I want to first point out that I have tried to avoid taking a stand in this paper on how the abstract truth conditions I assign to counterfactuals are related to their semantic (or assertoric) contents in context. One approach would be to let backtracking and non-backtracking interpretations of the same counterfactual express distinct propositions. This implementation would predict that you and Beth may both assert true propositions. However, an alternative approach would be to let the time and salient proposition parameters of a counterfactual be features of the index of evaluation (as in a two-dimensional Kaplanian framework) that are initialized by a context of assessment (as in Lasersohn 2005, MacFarlane 2014). That implementation would predict that the propositions asserted by you and Beth are such that, relative to any context of assessment, at least one of them must be false. I officially want to take no stand on this issue in this paper. However, in what follows, I offer some reason to think that disputes involving backtracking/non-backtracking interpretations are more complicated than the objection above lets on.

Consider again the scenario from the outset: Smith is poised to jumped on the top of a 20 story building with no net beneath him. Now compare the following two discourses:

(33) Discourse 1

a. **Avon**: If Smith had jumped, he would have died.

b. **Beth**: No. Smith was rational, had no wish to die, could see below him, and knew that jumping without a net would kill him. Therefore, had Smith jumped, there would already have been a net below him to catch him safely. Hence, if Smith had jumped, he would have lived.

(34) Discourse 2:

[54] Thanks to an anonymous reviewer for raising this concern.

a. **Joe**: If Smith had jumped, he would have died.
b. **Sue**: No. If Smith had jumped, he would have lived, since he would have landed in the net beneath him.

There are several intuitive differences between these discourses. For instance, Joe could appropriately correct Sue by saying 'But there was no net beneath him' and it seems Sue ought to retract her claim in light of learning that fact. On the other hand, although Avon could attempt to change the context back to favoring a non-backtracking interpretation of (1) by emphasizing that there was no net beneath Smith, in light of learning this fact (which Beth presumably already knows), Beth would not thereby have to retract her claim. Another intuitive difference is that Joe and Sue seem to be disagreeing over a common subject matter (what would have happened if Smith had jumped), whereas Avon and Beth are not. Granted, the subject matter of Avon and Beth's disagreement is not obvious, but this is still a difference between the discourses. Finally, it seems correct to say that Avon and Beth are merely talking past one another, and in particular that both may speak truly. By contrast, the intuition regarding Joe and Sue is that one of their claims must be false.

These intuitions favor handling backtracking/non-backtracking interpretations of counterfactuals as expressing distinct propositions. But then what about the intuition that in both cases denial (saying 'No') is linguistically appropriate? This is unlike paradigmatic cases of 'talking past' involving automatic indexicals, as in the following example from Lasersohn 2005, p. 647:

(35) a. Kara: I am a doctor.
b. Tim: #No, I am not a doctor.
c. Tom: No, you are not a doctor.

We might be tempted to generalize from this example that, in discourses licensing denial, the two parties genuinely disagree in the sense of making claims about a common subject matter, one of which must be false. After all, this is what is lacking in Kara/Tim's discourse that is not lacking in Kara/Tom's. If this generalization were correct, then the difference between backtracking/non-backtracking interpretations should not result in those different interpretations expressing distinct propositions. However, this generalization is mistaken. Consider cases like the following (cf. Horn 1985, Sundell 2011):

(36) A: Smith ate some of the cookies.

B: No, Smith ate all of the cookies.

(37) A: Burgers come with chips.
B: No they come with french fries.

Denial is licensed in both of these examples but in neither is it the case that at least one of the claims must be false. Therefore, there is room to think that denials are sometimes (though not always, as exhibited by Kara/Tim) licensed in exchanges where the two parties assert compatible propositions. Exactly what conditions must be in place for this to happen is a topic that goes far beyond this paper.[55]

5.4 Backtracking indicatives?

Objection: *There seem to be backtracking indicatives. This is a problem for your view because it predicts that backtracking interpretations arise only when a conditional has an earlier than default conditional time, and this is implausible for indicative conditionals since they are not past tensed.*

An alleged example of a backtracking interpretation of an indicative conditional comes from Gibbard's 'Sly Pete' case (cf. Gibbard 1981, Bennett 2003, DeRose 2010). Here is a version of that case:

> Poker Sly Pete is playing poker against Mr. Stone. Pete has two associates helping him cheat. Zack sees Stone's hand and signals its contents to Pete, and receives Pete's sign that he got the message. Jack sees both hands and sees that Pete has the losing hand. Before Pete makes his move, Stone gets suspicious and removes everyone else from the room.

After the hand is played, but before the results revealed, it seems that Zack will accept and Jack will reject (respectively):

(38) If Pete called, he won.

It seems plausible that there are two distinct interpretations of (38): one on which it is true (Zack's interpretation, $\mathsf{P} \rightarrow_Z \mathsf{W}$) and one on which it is false (Jack's interpretation,

[55] For some concrete proposals about this issue, see Sundell 2011, Plunkett & Sundell 2013, Khoo 2015a.

$\mathsf{P} \rightarrow_J \mathsf{W}$).[56] Jack's interpretation, $\mathsf{P} \rightarrow_J \mathsf{W}$, is false because its domain entails that Pete had the losing hand (though it does not entail that Pete called iff he had the winning hand). Zack's interpretation, $\mathsf{P} \rightarrow_Z \mathsf{W}$, is true because its domain entails that Pete called iff he had the winning hand (though it does not entail that Pete had the losing hand). As such, Zack's interpretation is similar in many respects to a backtracking interpretation of a counterfactual. He reasons as follows: if Pete called, then he had to have had the better hand (since he knew both hands and wanted to win), and so he won. Thus, it seems there are backtracking interpretations of indicatives.

Reply: I am happy to grant that Jack and Zack entertain distinct interpretations of (38), and I am also happy to grant that Zack's interpretation bears many of the features of backtracking interpretations of counterfactuals. However, there is still an important difference between the context-dependence exhibited by the indicative (38) and the context-dependence exhibited by the counterfactual (2), which suggests that we ought to embrace different accounts for each. (2) is by default non-backtracking, though this interpretation may be overruled by asserting and accepting some relevant backward counterfactual. By contrast, (38) is not by default interpreted as 'non-backtracking' (the interpretation of it that Jack rejects).[57] Rather, it seems that anyone in Zack's epistemic situation should interpret (38) on its 'backtracking' interpretation and thus accept it, even in the absence of some extra contextual preamble (and likewise for Jack and the 'non-backtracking' interpretation of (38)). Thus, unlike (2), the two interpretations of (38) brought out by Zack and Jack's epistemic situations seem to be on a par. Thus, the analog of (B) does not seem to hold for 'backtracking' interpretations of indicatives. Since (B) is explained by my theory's appeal to historical modality and since it is compatible with my theory that indicatives do not quantify over historically possible worlds, this feature of indicative 'backtrackers' is compatible with my theory.

One way of accounting for this difference between (2) and (38) is to adopt the theory I propose about the former and then adopt a theory of the latter on which it is an epistemic

[56] Although this is controversial—see Gibbard 1981, Stalnaker 1984, Bennett 2003, Williams 2008. I will grant the assumption for now.

[57] One might think that it is false merely on the grounds that Pete has the losing hand. However, as DeRose 2010 points out, Zack may think it very likely that Pete has the losing hand (just stipulate that Mr. Stone's hand is very good), and still have every reason to accept (38). Yet if Pete having the losing hand were in fact sufficient to falsify (38), then Zack thinking it very likely that Pete has the losing hand should lead him to think it very likely that (38) is false, in which case he would have good reason reject it.

conditional—hence about what some relevant agent knows or presupposes (cf. Ramsey 1931, Stalnaker 1975, Warmbrod 1983, Weatherson 2001, Williams 2008). On such a theory, the truth of an indicative conditional is relative to an information state. In evaluating (38) with respect to Zack's information, we add its antecedent to that information, while also adding other propositions about how its antecedent must have come to be given its addition to what he knows. (38) is true relative to Zack's information state because the modified information state (with the supposition of the antecedent) entails that Pete won. We may call this sort of reasoning 'backtracking,' but it isn't the same thing as what happens when a counterfactual has a backtracking interpretation. Rather, it is just a byproduct of what normally happens when we evaluate an indicative conditional according to the 'Ramsey test'—we suppose its antecedent, making the requisite changes to add it to our information state. Since both Jack and Zack engage in this process in evaluating (38), and since their two information states are on a par (both are ignorant of some relevant fact—Zack that Pete has the losing hand, and Jack that Pete knows Stone's hand), there is no asymmetry between the two interpretations. Thus, we predict this difference between (2) and (38).

5.5 Underspecified antecedents

Objector: *Consider the following scenario (from Bennett 2003: 219-220, who attributes it to Nute 1980):*

> `Thieves`. Unbeknownst to John, two thieves, Slim and Tim, were vying for his jacket at the restaurant. At 1:00pm, when John was in the bathroom, Slim made a move for the jacket, but was thwarted by the server, who happened by the table. At 1:05pm, when John was paying the check, Tim made his move, but was unable to nab it thanks again to the server, who bumped into him.

(39) If John's jacket had been stolen from the restaurant, it would now be in the possession of Tim.

Intuitively, (39) strikes us as unassertable given the information in `Thieves`*. But according to* **Default***, later admissible counterfactual times are better candidates for the default counterfactual time. If that is so, then we seem to predict, incorrectly, that (39) should be true and known (and hence assertable).*

Reply: This objection gives me the opportunity to discuss the important issue of antecedent underdetermination, an issue already raised in footnote 33 but set aside until

now. Probably all natural language counterfactual antecedents are underdetermined—that is, they do not specify which of several possible ways they are to be (hypothetically) realized. For instance, the antecedent of (39) is underdetermined: one way for John's coat to have been stolen is for Tim to have stolen John's coat, and another is for Slim to have stolen John's coat. Now that we have the notion of antecedent underdetermination on the table, to see that the issue raised by the `Thieves` example is one about antecedent underdetermination, notice that the same point can be made without distinguishing the times at which Tim and Slim make their moves:

> `Thieves-2`. Unbeknownst to John, two thieves, Slim and Tim, were vying for his jacket at the restaurant. Both planned to steal John's jacket whenever he was off-guard, but neither got the chance.

Just as in `Thieves`, (39) seems unassertable, given the information stated in `Thieves-2`. Thus, it seems that the reason (39) is unassertable should be the same in both scenarios. But since there is nothing about time mentioned in `Thieves-2`, it is unlikely that the former example is a problem about counterfactual time, in particular.

So why do we judge (39) unassertable in these contexts? An offhand diagnosis is that in these contexts there are two salient ways of realizing (39)'s antecedent: Slim stealing John's coat and Tim stealing John's coat. Furthermore, both ways are possible, given the background information. That is, it is possible that had John's jacket been stolen, it would have been stolen by Slim; and it is possible that had John's jacket been stolen, it would have been stolen by Tim. Hence, (39) is unassertable because it is not known, given the information in the case.

To incorporate this insight into our semantics, I propose to think of ways of realizing a proposition A as the truthmakers for A (in a vein similar to Fine 2012a,b, although my theory will make a very different use of this idea). For our purposes, we may simply hold that a truthmaker for A at w is a proposition that is true at w and entails A. Although I won't offer an account of what makes a truthmaker for a proposition *salient* in a context, it seems plausible that in `Thieves` and `Thieves-2`, the two c-salient truthmakers for (39)'s antecedent are that it is stolen by Tim and that it is stolen by by Slim, since in both contexts it is explicitly stated that both Tim and Slim in fact attempted to steal John's coat. Finally, we modify our account of **Possible antecedent** to make use of this extra machinery by ensuring that the resulting counterfactual domain is compatible with each of the c-salient truthmakers for A:

Possible antecedent*: Counterfactual time t must be such that, for each c-salient truthmaker for A, A^{n}: $D_c(\mathsf{A}^{\mathsf{n}}, w, t) \neq \emptyset$.

With these modifications to our theory in place, let us return to `Thieves`. Since both that Slim stole John's jacket and that Tim stole John's jacket are salient truthmakers in that context for (39)'s antecedent, by **Possible antecedent***, the counterfactual time for (39) must be before the times those propositions are about. Hence, we predict that both of the following are possible, given the information in `Thieves`:

(40) a. If John's jacket had been stolen from the restaurant, it would have been stolen by Tim.

b. If John's jacket had been stolen from the restaurant, it would have been stolen by Slim.

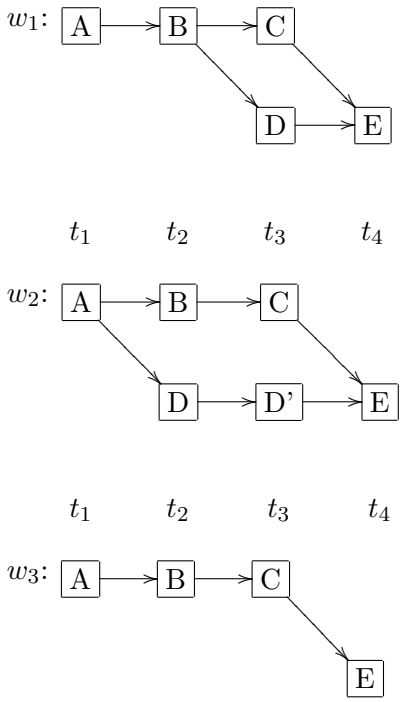
Therefore, since (39) is known only if it is known which of (40-a)/(40-b) is the case, (39) is predicted to be unknown and hence unassertable.[58]

Appendix

Counterfactuals and INFORMATIVITY

I here explore whether INFORMATIVITY may constrain the default interpretation of counterfactuals. Let '$\mathsf{A} \boxtimes\!\!\rightarrow_{\mathsf{t}} \mathsf{C}$' denote the proposition expressed by $A \boxtimes\!\!\rightarrow C$ as interpreted with counterfactual time t (that is, $\{w : D_c(\mathsf{A}, w, t) \subseteq \mathsf{C}\}$). My first observation is that, if the relevant causal sufficiencies and facts differ between worlds, then for any times $t < t'$, neither $\mathsf{A} \boxtimes\!\!\rightarrow_{\mathsf{t}} \mathsf{C}$ nor $\mathsf{A} \boxtimes\!\!\rightarrow_{\mathsf{t}'} \mathsf{C}$ will strictly entail the other. To illustrate why, consider the following simple example involving the following three worlds:

[58]Thanks to audiences at MIT and at the Bellingham Summer Philosophy Conference for helpful feedback on earlier drafts. In particular, thanks to Simona Aimar, Andy Egan, Tom Dougherty, Irene Heim, Aron Hirsch, Brendan Jackson, Josh Knobe, Rose Lenehan, Karen Lewis, Matt Mandelkern, Vann McGee, Sarah Moss, Dilip Ninan, David Plunkett, Bernhard Salow, Raul Saucedo, Jonathan Schaffer, Joshua Schechter, Miriam Schoenfield, Ginger Schultheis, Brad Skow, Bob Stalnaker, Eric Swanson, Zoltán Gendler Szabó, and Steve Yablo for insightful comments.



Now consider the following counterfactual, $\neg C \mathrel{\Box\!\!\rightarrow} E$, and assume that D is salient. Given our semantics, it follows that $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_2}} \mathsf{E}$ true at w_1, but $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_1}} \mathsf{E}$ is false at w_1. Therefore, $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_2}} \mathsf{E}$ does not entail $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_1}} \mathsf{E}$.

- **Why $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_2}} \mathsf{E}$ is true at w_1.** $D_c(\overline{\mathsf{C}}, w_1, t_2)$ entails A and B (by HISTORICAL). Furthermore, since D is caused in the same way at D-worlds in $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_2}$ as it is at w_1 (caused by B), by HINDSIGHT, $D_c(\overline{\mathsf{C}}, w_1, t_2) \models \mathsf{D}$. Finally, since $\mathsf{D} \supset \mathsf{E}$ is a causal sufficiency of w_1 about times later than t_2, $D_c(\overline{\mathsf{C}}, w_1, t_2) \models \mathsf{D} \supset \mathsf{E}$ (by CAUSAL). Hence, by closure, $D_c(\overline{\mathsf{C}}, w_1, t_2) \models \mathsf{E}$.

- **Why $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_1}} \mathsf{E}$ is false at w_1.** $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_1} \models \mathsf{B} \supset \mathsf{C}$ (since $\mathsf{B} \supset \mathsf{C}$ is a causal sufficiency of w_1 about times after t_1). However, since $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_1} \models \overline{\mathsf{C}}$, by closure, $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_1} \models \overline{\mathsf{B}}$. Thus, D is not caused in the same way at D-worlds in $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_1}$ as it is at w_1. At w_1, D is caused by B, but B does not cause D at any D-world in $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_1}$, since B is false at every one of those worlds. Hence, HINDSIGHT does not ensure that $D_c(\overline{\mathsf{C}}, w_1, t_1) \models \mathsf{D}$. But notice that nothing else in the theory ensures this, so $D_c(\overline{\mathsf{C}}, w_1, t_1) \not\models \mathsf{D}$. And therefore, since the same reasoning applies to E, $D_c(\overline{\mathsf{C}}, w_1, t_1) \not\models \mathsf{E}$.

Next, $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_1}} \mathsf{E}$ is true at w_2 but $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_2}} \mathsf{E}$ is false at w_3. So, $\overline{\mathsf{C}} \mathrel{\Box\!\!\rightarrow_{\mathsf{t}_1}} \mathsf{E}$ does not entail

$\overline{\mathsf{C}} \,\Box\!\!\rightarrow_{\mathsf{t}_2} \mathsf{E}$.

- **Why $\overline{\mathsf{C}} \,\Box\!\!\rightarrow_{t_1} \mathsf{E}$ is true at** w_2. $D_c(\overline{\mathsf{C}}, w_2, t_1)$ entails A (by HISTORICAL). However, it also entails D', by HINDSIGHT. This is because D' is caused in the same way at D'-worlds in $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_1}$ as it is at w_2 (being caused by D which is caused by A). And by CAUSAL, $D_c(\overline{\mathsf{C}}, w_2, t_1) \models \mathsf{D}' \supset \mathsf{E}$. Therefore, $D_c(\overline{\mathsf{C}}, w_2, t_1) \models \mathsf{E}$.

- **Why $\overline{\mathsf{C}} \,\Box\!\!\rightarrow_{\mathsf{t}_2} \mathsf{E}$ is false at** w_3. Note that the only way $D_c(\overline{\mathsf{C}}, w_3, t_2)$ would entail E is by HINDSIGHT, since no causal sufficiencies will entailed by $D_c(\overline{\mathsf{C}}, w_3, t_2)$ will ensure that it entails E. But then notice that E is not caused in the same way at E-worlds in $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_2}$ as it is at w_3. At w_3, E is caused by C which is caused by B. At no E-world in $\mathsf{HS}^{\overline{\mathsf{C}}}_{\mathsf{t}_2}$ is it caused by C. So, HINDSIGHT does not ensure that $D_c(\overline{\mathsf{C}}, w_3, t_2) \models \mathsf{E}$. But then nothing ensures this. Thus, $D_c(\overline{\mathsf{C}}, w_3, t_2) \not\models \mathsf{E}$.

Therefore, since it is generally the case that the worlds in $\mathcal{W}$ (or compatible with what is presupposed, or perhaps taken to be known) differ with respect to the relevant causal sufficiencies and facts, it will almost always be the case that neither temporal interpretation is more informative in the relevant sense (of strictly entailing the less informative interpretation), and thus INFORMATIVITY will generally provide no additional constraint on interpretation.

References

Abusch, Dorit. 1997. Sequence of tense and temporal de re. *Linguistics and Philosophy*, **20**, 1–50.

Abusch, Dorit. 1998. Generalizing Tense Semantics for Future Contexts. *Pages 13–35 of:* Rothstein, Susan (ed), *Events and Grammar*. Dordrecht: Kluwer Academic Publishers.

Ahmed, Arif. 2010. Out of the Closet. *Analysis*, 1–9.

Ahmed, Arif. 2011. Walters on Conjunction Conditionalization. *Proceedings of the Aristotelian Society*, 115–122.

Arregui, Ana. 2005a. *Layering Modalities: The Case of Backtracking Counterfactuals*. ms.

Arregui, Ana. 2005b. *On the Accessibility of Possible Worlds: the Role of Tense and Aspect*. Ph.D. thesis, University of Massachusetts at Amherst.

Arregui, Ana. 2007. When Aspect Matters: The Case of *Would* Conditionals. *Natural Language Semantics*, **15**, 221–264.

Arregui, Ana. 2009. On Similarity in Counterfactuals. *Linguistics and Philosophy*, **32**, 245–278.

Barker, Stephen. 1991. *Even, Still*, and Counterfactuals. *Linguistics and Philosophy*, **14**, 1–38.

Barker, Stephen. 1994. The Consequent-Entailment Problem for *Even If*. *Linguistics and Philosophy*, **17**, 249–260.

Barker, Stephen J. 1998. Predetermination and Tense Probabilism. *Analysis*, **58**(4), 290–296.

Bennett, Jonathan. 1982. Even If. *Linguistics and Philosophy*, **5**, 403–418.

Bennett, Jonathan. 1984. Counterfactuals and Temporal Direction. *The Philosophical Review*, **93**(1), 57–91.

Bennett, Jonathan. 2003. *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Berman, Stephen. 1987. Situation-Based Semantics for Adverbs of Quantification. *In:* Blevins, J., & Vainikka, A. (eds), *UMOP*, vol. 12. Amherst: GLSA.

Briggs, Rachael. 2012. Interventionist Counterfactuals. *Philosophical Studies*, **160**, 139–166.

Carston, Robyn. 2002. *Thoughts and Utterances: the Pragmatics of Explicit Communication*. Oxford: Blackwell.

Chisholm, Roderick. 1955. Law Statements and Counterfactual Inference. *Analysis*, **15**, 97–105.

DeRose, Keith. 1994. Lewis on 'Might' and 'Would' Counterfactual Conditionals. *Canadian Journal of Philosophy*, **24**(3), 413–418.

DeRose, Keith. 1999. Can It Be That It Would Have Been Even Though It Might Not Have Been? *Philosophical Perspectives*, **13**, 385–413.

DeRose, Keith. 2010. The Conditionals of Deliberation. *Mind*, **119**(473), 1–42.

Dowell, Janice. 2012. Contextualist Solutions to Three Puzzles about Practical Conditionals. *Pages 271–303 of:* Shafer-Landau, Russ (ed), *Oxford Studies in Metaethics*, vol. 7. Oxford: Oxford University Press.

Dowell, Janice. 2013. Flexible Contextualism about Deontic Modals: A Puzzle about Information-Sensitivity. *Inquiry*, **56**(2-3), 149–178.

Downing, P.B. 1959. Subjunctive Conditionals, Time Order, and Causation. *Meeting of the Aristotelian Society*, **59**, 125–140.

Dudman, V. H. 1983. Tense and Time in English Verb Clusters of the Primary Pattern. *Australasian Journal of Linguistics*, **3**, 25–44.

Dudman, V. H. 1984. Parsing 'If'-Sentences. *Analysis*, **44**(4), 145–153.

Dudman, V. H. 1988. Indicative and Subjunctive. *Analysis*, **48**(3), 113–122.

Edgington, Dorothy. 1995. On Conditionals. *Mind*, **104**, 235–329.

Edgington, Dorothy. 2004. Counterfactuals and the Benefit of Hindsight. *Pages 12–28 of: Cause and Chance: Causation in an Indeterministic World.* New York: Routledge.

Elga, Adam. 2001. Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science*, **68**(3), S313–S324.

Fine, Kit. 1975. Critical Notice: *Counterfactuals. Mind*, **84**(335), 451–458.

Fine, Kit. 2012a. Counterfactuals Without Possible Worlds. *Journal of Philosophy*, **109**(3), 221–246.

Fine, Kit. 2012b. A Difficulty for the Possible Worlds Analysis of Counterfactuals. *Synthese*, **189**(1), 29–57.

von Fintel, Kai. 1997. The Presupposition of Subjunctive Conditionals. *In:* Percus, Orin, & Sauerland, Uri (eds), *MIT Working Papers in Linguistics*, vol. 25.

von Fintel, Kai. 2001. Counterfactuals in a Dynamic Context. *Pages 123–152 of:* Kenstowicz, Michael (ed), *Ken Hale: a Life in Language.* Cambridge: MIT Press.

von Fintel, Kai. 2004. A Minimal Theory of Adverbial Quantification. *Pages 153–193 of:* Partee, Barbara, & Kamp, Hans (eds), *Context Dependence in the Analysis of Lingusitic Meaning.* Amsterdam: Elsevier.

von Fintel, Kai, & Iatridou, Sabine. 2008. How to Say *Ought* in Foreign: the Composition of Weak Necessity Modals. *Pages 115–141 of:* Guéron, Jacqueline, & Lecarme, Jacqueline (eds), *Time and Modality.* Springer.

Gerstenberg, Tobias, Bechlivanidis, Christos, & Lagnado, David A. 2013. Back on Track: Backtracking in Counterfactual Reasoning. *In:* Knauff, M., Pauen, M., Sebanz, N., & Wachsmuth, I. (eds), *Proceedings of the 35th Annual Conference of the Cognitive Science Society.*

Gibbard, Allan. 1981. Two Recent Theories of Conditionals. *Pages 211–247 of:* Harper, William L., Stalnaker, Robert, & Pearce, Glenn (eds), *Ifs*. Dordrecht: Reidel.

Gillies, Anthony. 2007. Counterfactual Scorekeeping. *Linguistics and Philosophy*, **30**, 329–360.

Glanzberg, Michael. 2007. Context, Content, and Relativism. *Philosophical Studies*, **136**, 1–29.

Glanzberg, Michael. 2015. *Not All Contextual Parameters are Alike*. ms.

Goodman, Nelson. 1947. The Problem of Counterfactual Conditionals. *Journal of Philosophy*, **44**(5), 113–128.

Hart, H. 1951. A Logician's Fairy Tale. *The Philosophical Review*, **60**, 198–212.

Hiddleston, Eric. 2005. A Causal Theory of Counterfactuals. *Nous*, **39**(4), 632–657.

Horn, Laurence. 1985. Metalinguistic Negation and Pragmatic Ambiguity. *Language*, **61**(1), 121–174.

Iatridou, Sabine. 2000. The Grammatical Features of Counterfactuality. *Linguistic Inquiry*, **31**, 231–270.

Ippolito, Michela. 2003. Presuppositions and Implicatures in Counterfactuals. *Natural Language Semantics*, **11**, 145–186.

Ippolito, Michela. 2006. Semantic Composition and Presupposition Projection in Subjunctive Conditionals. *Linguistics and Philosophy*, **29**, 631–672.

Ippolito, Michela. 2013a. Counterfactuals and Conditional Questions Under Discussion. *Pages 194–211 of:* Snider, Todd (ed), *Proceedings of SALT 23*.

Ippolito, Michela. 2013b. *Subjunctive Conditionals: a Linguistic Analysis*. Linguistic Inquiry Monograph Series. Cambridge: MIT Press.

Isard, Steven. 1974. What Would You Have Done If .. *Theoretical Linguistics*, **1**, 233–255.

Jackson, Frank. 1977. A Causal Theory of Counterfactuals. *Australasian Journal of Philosophy*, **55**, 3–21.

Kaplan, David. 1989. Demonstratives. *Pages 481–563 of:* Almog, Joseph, Perry, John, & Wettstein, Howard (eds), *Themes from Kaplan*. Oxford: Oxford University Press.

Kaufmann, Stefan. 2005. Conditional Predictions. *Linguistics and Philosophy*, **28**, 181–231.

Khoo, Justin. 2015a. Modal Disagreements. *Inquiry*, **58**(5), 511–534.

Khoo, Justin. 2015b. On Indicative and Subjunctive Conditionals. *Philosophers' Imprint*, **15**(32), 1–40.

King, Jeffrey C. 2013a. The Metasemantics of Contextual Sensitivity. *In:* Burgess, A., & Sherman, B. (eds), *New Essays in Metasemantics*. Oxford University Press.

King, Jeffrey C. 2013b. Supplementives, the Coordination Account, and Conflicting Intentions. *Philosophical Perspectives*, **27**, 288–311.

King, Jeffrey C. 2014. Speaker Intentions in Context. *Nous*, **48**(2), 219–237.

Kment, Boris. 2006. Counterfactuals and Explanation. *Mind*, **115**(458), 261–309.

Kment, Boris. 2014. *Modality and Metaphysical Explanation*. Oxford: Oxford University Press.

Kratzer, Angelika. 1989. An Investigation of the Lumps of Thought. *Linguistics and Philosophy*, **12**(5), 607–653.

Kratzer, Angelika. 2012. *Collected Papers on Modals and Conditionals*. Oxford: Oxford University Press.

Lange, Marc. 2000. *Natural Laws in Scientific Practice*. Oxford: Oxford University Press.

Lange, Marc. 2009. *Laws and Lawmakers*. New York: Oxford University Press.

Lasersohn, Peter. 2005. Context Dependence, Disagreement, and Predicates of Personal Taste. *Linguistics and Philosophy*, **28**(6), 643–686.

Levinson, Stephen. 2000. *Presumptive Meanings: The Theory of the Generalized Conversational Implicature*. Cambridge: MIT Press.

Lewis, David. 1973a. *Counterfactuals*. Oxford: Blackwell.

Lewis, David. 1973b. Counterfactuals and Comparative Possibility. *Journal of Philosophical Logic*, **2**, 418–446.

Lewis, David. 1979a. Counterfactual Dependence and Time's Arrow. *Nous*, **13**, 455–476.

Lewis, David. 1979b. Scorekeeping in a Language Game. *Journal of Philosophical Logic*, **8**, 339–59.

Lewis, David. 1981. Ordering Semantics and Premise Semantics for Counterfactuals. *Journal of Philosophical Logic*, **10**, 217–234.

Lewis, David. 1986. Causation. *Pages 159–172 of: Philosophical Papers Vol 1*. New York: Oxford University Press.

Lewis, David. 2000. Causation as Influence. *Journal of Philosophy*, **97**, 182–197.

Lycan, William. 1991. *Even* and *Even If*. *Linguistics and Philosophy*, **14**, 115–150.

Lycan, William. 2001. *Real Conditionals*. Oxford: Oxford University Press.

Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.

MacFarlane, John. 2014. *Assessment Sensitivity: Relative Truth and its Applications*. Oxford: Oxford University Press.

Maudlin, Tim. 2007. *The Metaphysics Within Physics*. New York: Oxford University Press.

Noordhof, Paul. 2005. Morgenbesser's Coin, Counterfactuals, and Independence. *Analysis*, **65**(3), 261–263.

Nute, Donald. 1980. *Topics in Conditional Logic*. Dordrecht: Reidel.

Ogihara, Toshiyuki. 1996. *Tense, Attitudes, and Scope*. Dordrecht: Kluwer Academic Publishers.

Palmer, Frank. 1986. *Mood and Modality*. Cambridge: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Perry, John. 1997. Indexicals and Demonstratives. *Pages 586–612 of:* Hale, Bob, & Wright, Crispin (eds), *A Companion to Philosophy of Language*. Oxford: Blackwell.

Phillips, Ian B. 2007. Morgenbesser Cases and Closet Determinism. *Analysis*, **67**, 42–49.

Phillips, Ian B. 2011. Stuck in the Closet: a Reply to Ahmed. *Analysis*, **71**, 86–91.

Placek, Tomasz, & Müller, Thomas. 2007. Counterfactuals and Historical Possibility. *Synthese*, **154**, 173–197.

Plunkett, David, & Sundell, Tim. 2013. Disagreement and the Semantics of Normative and Evaluative Terms. *Philosophers' Imprint*, **13**(23), 1–37.

Ramsey, Frank P. 1931. *The Foundations of Mathematics and other Logical Essays*. London: Kegan Paul, Trench, Trubner & Co.

Reichenbach, Hans. 1947. *Elements of Symbolic Logic*. New York: Collier-Macmillan.

Richard, Mark. 2004. Contextualism and Relativism. *Philosophical Studies*, **119**, 215–242.

Roberts, Craige. 2012a. Information Structure: Afterword. *Semantics & Pragmatics*, **5**(7), 1–19.

Roberts, Craige. 2012b. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. *Semantics & Pragmatics*, **5**(6), 1–69.

Rosen, Gideon. 2014. *What is Normative Necessity?* ms.

Schaffer, Jonathan. 2004a. Counterfactuals, Causal Independence, and Conceptual Circularity. *Analysis*, **64**(4), 299–309.

Schaffer, Jonathan. 2004b. From Contextualism to Contrastivism. *Philosophical Studies*, **119**, 73–103.

Schulz, Katrin. 2014. Fake Tense in Conditional Sentences: A Modal Approach. *Natural Language Semantics*.

Slote, Michael. 1978. Time in Counterfactuals. *The Philosophical Review*, **87**(1), 3–27.

Sperber, Dan, & Wilson, Deirdre. 1986. *Relevance*. Oxford: Blackwell.

Stalnaker, Robert. 1968. A Theory of Conditionals. *Pages 98–112 of:* Rescher, N. (ed), *Studies in Logical Theory*. Oxford: Oxford University Press.

Stalnaker, Robert. 1975. Indicative Conditionals. *Philosophia*, **5**, 269–86.

Stalnaker, Robert. 1980. A Defense of Conditional Excluded Middle. *In:* Harper, William L., Pearce, Glenn, & Stalnaker, Robert (eds), *Ifs*. Dordrecht: Reidel.

Stalnaker, Robert. 1984. *Inquiry*. MIT Press.

Starr, William B. 2013. A Uniform Theory of Conditionals. *Journal of Philosophical Logic*.

Strawson, P.F. 1952. *Introduction to Logical Theory*. London: Methuen.

Sundell, Timothy. 2011. Disagreements About Taste. *Philosophical Studies*, **155**, 267–288.

Swanson, Eric. 2011. Conditional Excluded Middle Without the Limit Assumption. *Philosophy and Phenomenological Research*.

Taylor, Kenneth. 2001. Sex, Breakfast, and Descriptus Interruptus. *Synthese*, **128**(45-61).

Tedeschi, Philip. 1981. Some Evidence for a Branching-Futures Semantic Model. *Pages 239–269 of:* Tedeschi, P., & Zaenen, A. (eds), *Syntax and Semantics: Tense and Aspect*, vol. 14. New York: Academic Press.

Thomason, Richmond. 1985. Note on Tense and Subjunctive Conditionals. *Philosophy of Science*, **52**(1), 151–153.

Thomason, Richmond, & Gupta, Anil. 1980. A Theory of Conditionals in the Context of Branching Time. *The Philosophical Review*, **89**(1), 65–90.

Tichý, Pavel. 1976. A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals. *Philosophical Studies*, **29**, 271–273.

Tooley, Michael. 2002. Backward Causation and the Stalnaker-Lewis Approach to Counterfactuals. *Analysis*, **62**(3), 191–197.

Tooley, Michael. 2003. The Stalnaker-Lewis Approach to Counterfactuals. *Journal of Philosophy*, **100**(7), 371–377.

Walters, Lee. 2009. Morgenbesser's Coin and Counterfactuals with True Components. *Proceedings of the Aristotelian Society*, **109**(3), 365–379.

Warmbrod, Ken. 1983. Epistemic Conditionals. *Pacific Philosophical Quarterly*, **64**, 249–265.

Wasserman, Ryan. 2006. The Future Similarity Objection Revisited. *Synthese*, **150**, 57–67.

Weatherson, Brian. 2001. Indicative and Subjunctive Conditionals. *The Philosophical Quarterly*, **51**(203), 200–216.

Williams, J. Robert G. 2008. Conversation and Conditionals. *Philosophical Studies*, **138**, 211–223.

Williamson, Timothy. 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

Wilson, Deirdre, & Sperber, Dan. 2012. *Meaning and Relevance*. Cambridge: Cambridge University Press.

Won, Chiwook. 2009. Morgenbesser's Coin, Counterfactuals, and Causal Versus Probabilistic Independence. *Erkenntnis*, **71**, 345–354.

Zuber, Richard. 1977. Decomposition of Factives. *Studies in Language*, **1**(4), 407–421.