The semantics and logic of counterfactuals[1]

Paolo Santorio ◦ University of Maryland, College Park

1 Introduction

Suppose that I'm holding a dry match. I don't strike it. As a result, the match does not light. Now consider:

(1) If I had struck the match, it would have lit.

(2) If I had struck the match, it would have been wet.

(1) and (2) are **counterfactuals**, namely—on a first pass—conditional sentences involving contrary-to-fact suppositions. (1) is true, provided that we grant some assumptions: I strike the match with enough force, there is no wind, etc. (2) is false. How should we explain this asymmetry?

Intuitively, we want to say something like the following. When supposing that I struck the match, we 'hold fixed' certain salient facts. For example, we hold fixed the fact that the match is dry. Conversely, there are equally salient facts that we do not hold fixed: for example, the fact that the match didn't light. What facts we hold fixed determines our judgments about counterfactuals like (1) and (2). The supposition that the match is struck, together with propositions that describe the facts that are held fixed, entail that the match lights up, but not that the match is wet. The truth-values of counterfactuals like (1) and (2) track what happens in this kind of suppositional reasoning.

This is progress, but it pushes back the main question. In virtue of what do we hold fixed certain facts and not others when evaluating counterfactuals? Notice that this is not a matter of context dependence or speakers' choice. There just is no plausible context (barring abstruse background stories and science-fictional matches) where we hear (2) as true and (1) as false. Whatever generates the asymmetry between (1) and (2) seems to be hardwired in their semantics.

This problem was first pointed out by Nelson Goodman (1947, 1955). Goodman calls the assumptions that we hold fixed when evaluating counterfactuals 'cotenable'. For him, the problem posed by counterfactuals is the problem of characterizing cotenability.

[1]Thanks to Ilaria Canavotto, Fabrizio Cariani, Matt Mandelkern, and Malte Willer for helpful comments on earlier drafts of this paper.

Since Goodman, counterfactuals have been at the center of an extremely large body of work, both in philosophy and outside. In part, this is simply because of the intrinsic interest of the topic. Counterfactual statements and suppositions are an interesting and important part of human psychology. But in part, this is because counterfactuals have earned a central place in a vast array of debates. In philosophy, counterfactuals have been linked to theories of causation, laws of nature, mental content, and knowledge, just to mention a few examples. The reason is that counterfactuals are one of the main ways to state necessary connections between events—where the kind of necessity in play is not epistemic, but rather has to do with how things are in the world. For the same reason, counterfactuals have been the subject of extensive research in psychology: the psychology of counterfactuals has a large overlap with the psychology of causal reasoning.

Modern philosophical work on counterfactuals starts with the idea that Goodman's problem becomes more tractable if we split it into two.[2] The **logical subproblem** is the problem of specifying a general template for the meaning of counterfactuals, in a way that we can predict what inferences involving counterfactuals are good or bad. The literature has done this by using certain formal tools, like selection functions or relations of comparative similarity. The **similarity sub-problem** is the problem of linking the formal notions that are used to solve the first problem to notions that have intuitive content.

This article is an opinionated guide to how the literature has faced these two sub-problems and to the many open issues relating to them. It also makes forays in nearby territories: the connection between counterfactual morphology and counterfactual meaning, the relation between counterfactuals and probability, and the so-called causal models framework. Overall, I will highlight how, even though the literature has made real progress, many crucial questions are still open.

I proceed as follows. §2 discusses the domain of inquiry. §3 presents the classical semantics of Lewis, Stalnaker, and Kratzer, and §4 provides an overview of the main theories of similarity. §5 considers some recent issues on the logic of counterfactuals. §6 discusses counterfactual morphology, §7 the interaction between probability and counterfactuals, and §8 the causal models framework.

2 The domain of inquiry

What counts as counterfactual modality? A first-pass answer is: the modality that is standardly expressed in English via the modal auxiliary *would*, and via correspond-

[2]These two problems correspond, roughly at least, to the 'logical problem' and the 'pragmatic problem' distinguished by Stalnaker 1968.

ing morphology in other languages. One might object that we should give a more illuminating characterization, perhaps pointing to some semantic features. But a characterization of this sort is surprisingly difficult to give.

The label 'counterfactual' conveys the suggestion that the relevant kind of modality requires making a contrary-to-fact supposition. But it is well-known that this is incorrect. A number of felicitous *would*-conditionals have antecedents compatible with speakers' information. Here is a classic example from Anderson 1951. Suppose that a doctor is examining a patient and entertains the hypothesis that they have measles. The doctor observes, in support of this hypothesis:

(3) If the patient had measles, they would be showing exactly the symptoms that they have. Hence the patient might well have measles.

(3) is used to provide evidence for the hypothesis expressed by its antecedent. So the antecedent is compatible with the speakers' beliefs. In addition, as I point out in §6, some *would*-conditionals not only allow, but require that the antecedent be compatible with the speakers' information.

A second label frequently used is 'subjunctive', suggesting that what characterizes counterfactuals is a certain grammatical mood (and its semantic effects). This is also inaccurate. Languages where mood morphology is more clearly visible, such as Italian and French, show that subjunctive mood is not needed to express the relevant kind of modality. In these languages, counterfactuals are routinely expressed via a combination of past tense, indicative mood, and imperfective aspect. Here is an example from Italian:

(4) Se Maria era a Roma, ero più contento.
If Maria past-be-IMPFV-3rd in Rome, past-be-IMPFV-1st more happy
"If Maria was in Rome, I would be happier"

As a result, the cleanest strategy is the one we started from: we characterize the relevant class of modal expressions by appealing simply to *would*-morphology. I will say more about what this involves in §6. von Fintel & Iatridou 2022 propose the label 'X-marking', to signal the fact that the relevant conditionals involve 'extra' morphological features with respect to indicatives.[3] Here I will stick with 'counterfactuals', which is undoubtedly the most popular label in philosophy.

[3]von Fintel and Iatridou also point out that the relevant morphology appears outside of conditional constructions, for example in wish reports and in some deontic modals.

3 Comparative similarity semantics

This section provides an overview of the classical semantic theories proposed by Stalnaker, Lewis, and Kratzer. It focuses on the first of the two sub-problems mentioned in §1, i.e. the problem of specifying a general form for the truth-conditions of counterfactuals that is sufficient to fix a logic.

Before starting, one clarification. Philosophers have standardly taken the logical form of counterfactuals to involve a binary connective that links antecedent and consequent. This connective is often represented as '$\Box\!\!\rightarrow$' after Lewis. Conversely, theories in linguistic semantics tend to adopt the so-called restrictor analysis, after Kratzer (1986, 2012). On this view, all conditionals are modalized statements in which the *if*-clause works as a restrictor of the domain of quantification of the modal. In particular, counterfactuals involve two logical operators: the modal *would*, and *if*. *would* is a quantifier over possible worlds, and the *if*-clause restricts the domain of quantification to worlds that make the antecedent true. On this second view, the logical form of counterfactuals does not involve a unique operator that encodes a counterfactual meaning. This difference is substantial, but it won't matter much for the topics touched on in this essay. So I will adopt the binary connective analysis, with the understanding that everything I say can be restated in terms of the restrictor analysis.

3.1 Comparative similarity semantics

Virtually all modern semantics for counterfactuals are based on some version of a simple idea, which Stalnaker puts pithily as follows:

> Consider a possible world in which A is true, and which otherwise differs minimally from the actual world. "If A, then B" is true (false) just in case B is true (false) in that possible world. (1968, p. 102)

For illustration, consider again (1):

(1) If I had struck the match, it would have lit.

On Stalnaker's intuitive gloss, (1) is true just in case the match lights in the 'minimally different' world (or worlds) where I strike the match. The challenge, of course, is to capture in a precise way this intuitive notion of 'minimal difference'. To achieve this, different theorists develop different formal tools. Below, I survey some of the classic theories in the literature, including Stalnaker's, Lewis's, and Kratzer's. To make comparisons easier, I will cast all of of their accounts in terms of comparative similarity (I come back to premise semantics frameworks, which Kratzer uses, in

§3.3). Also, for the time being I stick to the assumption that counterfactuals have a classical truth conditional semantics.[4]

Our basic formal tool is a relation of comparative similarity (or closeness), represented as '$\preceq_w$'. $\preceq_w$ compares worlds with respect to their similarity to a benchmark world w: $\ulcorner w' \preceq_w w'' \urcorner$ says that w' is at least as similar (close) to w than w'' is. Comparative similarity is the formal counterpart of an informal notion of similarity, which I discuss in §4.

The exact way in which $\preceq_w$ figures in the truth conditions for counterfactuals varies across specific versions of the semantics. Here I present four versions of these truth conditions. They are progressively more complex, and logically weaker.

The first version relies on the assumption that, for each world w and each antecedent A, there is a single A-verifying world w' that is most similar, or closest, to w. With this assumption in the background, we can state the truth conditions of counterfactuals as follows:

(CS1) $\ulcorner$If A, would B$\urcorner$ is true at w iff the $\preceq_w$-closest A-world to w is a B-world

(CS1) is a restatement, in comparative similarity terms, of Stalnaker's single-world selection semantics (1968, 1981). As is evident, the semantics imposes fairly stringent requirements on comparative similarity. In particular, it requires that $\preceq_w$ induce a **linear order** on worlds (see Table 1). Graphically, this order can be represented very simply: worlds are put 'on a line'; there are no ties in which worlds count as closer or farther off from the world of evaluation (see Figure 1).

The second version assumes not that there is a single closest worlds, but allows that there might be a set of them.[5] Counterfactuals are universal quantifiers over this set.

---

[4]Rather than comparative similarity, Stalnaker's semantics uses *selection functions*. Informally, a selection function maps a world and a sentence to the closest antecedent world that makes that sentence true. More formally, selection functions are functions from a 'base' world and a proposition $s : W \times \mathcal{P}(W) \mapsto W$ that satisfy four conditions, listed below. (I use '$[\![\mathrm{A}]\!]$' to denote the proposition expressed by sentence A.)

i. if $[\![\mathrm{A}]\!]$ is non-empty, $s(w, \mathrm{A}) \in [\![\mathrm{A}]\!]$
(**Inclusion:** the selected world must make true the input sentence, if possible.)

ii. if $s(w, \mathrm{A}) = \lambda$, then $[\![\mathrm{A}]\!] = \varnothing$ (where λ is the absurd world, where every sentence is true)
(**Absurdity-as-last-resort**: λ is selected only if no possible world can be selected.)

iii. if $w \in [\![\mathrm{A}]\!]$, then $s(w, \mathrm{A}) = w$
(**Centering**: if the world of evaluation makes the input sentence true, it is the selected world.)

iv. for all A, A′: if $s(w, \mathrm{A}) \in [\![\mathrm{A}']\!]$ and $s(w, \mathrm{A}') \in [\![\mathrm{A}]\!]$, then $s(w, \mathrm{A}) = s(w, \mathrm{A}')$
(**Consistency of selection**: the selection must be consistent for all choice of input sentences.)

[5]The notion of a closest world to w is defined from $\preceq_w$ as follows: w' is among the closest A-worlds to w iff there is no A world w'' such that (i) $w'' \preceq_w w'$ and (ii) it is not the case that $w' \preceq_w w''$.

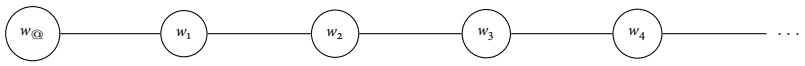


Figure 1: A linear order of worlds

(CS2) ⌜If A, would B⌝ is true at w iff, for every world w' that is among the $\preceq_w$-closest A-worlds to w, w' is a B-world

Differently from (CS1), (CS2) allows that different worlds might be tied for similarity. As a result, $\preceq_w$ induces a **total preorder** (or weak order) on worlds. To make things intuitive, it is useful to think of a total preorder as a linear order of sets of worlds, where worlds in the same set as tied for closeness. Graphically, this can be represented as in Figure 2.

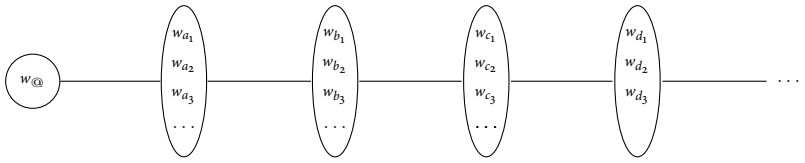


Figure 2: A total preorder of worlds

The semantics in (CS2) still presupposes that, for every world w, and for every counterfactual antecedent A, there is a set of worlds that are closest to w and that make A true. Famously, this assumption—the so-called **limit assumption**—is questioned by Lewis (1973, chapter 1). Lewis points out that there might be cases where we have an infinite sequence of antecedent-verifying worlds, where each world in the sequence is closer to the world of evaluation than its predecessor. Here is his example. Suppose that, in the actual world @, we have a printed line whose length is exactly one inch. Now considers a counterfactual starting with the antecedent *If the line had been longer than it is...*. What are the closest world to @ where this antecedent is true? According to Lewis, there are none. For *reductio*, suppose that we have a set of closest worlds S, in which the line has length n, where n is slightly longer than 1 inch. The problem is that, no matter how small n is, we can find worlds where the line is shorter than n, but is still longer than 1 inch. Those worlds are, according to Lewis, closer to @ than the worlds in the S, contrary to our assumption.

To make room for the failure of the limit assumption, we weaken the truth conditions in (CS2) as follows:

(CS3) ⌜If A, would B⌝ is true at w iff there is an A-world w'' such that, for every A-world w' that is at least as close to w as w'', w' is a B-world

Reflexivity.	For all w': $w' \preceq_w w'$
Transitivity.	For all w', w'', w''': if $w' \preceq_w w''$ and $w'' \preceq_w w'''$, then $w' \preceq_w w'''$
Antisymmetry.	For all w', w'': if $w' \preceq_w w''$ and $w'' \preceq_w w'$, then $w' = w''$
Strong connectedness.	For all w', w'': $w' \preceq_w w''$ or $w'' \preceq_w w'$
LINEAR ORDER	reflexive, transitive, antisymmetric, strongly connected
TOTAL PREORDER	reflexive, transitive, strongly connected
PREORDER	reflexive, transitive

Table 1: Formal features of the comparative similarity relation on various accounts.

These truth conditions require that, at some point, we are able to find an antecedent-verifying world such that every antecedent-verifying world that is at least as close as it is a consequent-verifying world. These truth conditions, which are Lewis's official truth conditions for counterfactuals, don't require that there be a set of closest antecedent-verifying worlds for all antecedents. (If, for some antecedent A, this set exists, the truth conditions in (CS3) reduce to the truth conditions in (CS2).)

Proponents of (CS3) still make a substantial assumption about similarity: all worlds are comparable with respect to similarity. For every two worlds w_1 and w_2, we have that w_1 is closer to w than w_2, w_2 is closer to w than w_1, or that w_1 and w_2 are equally close. Some question this assumption. It might be that $\preceq_w$ just fails to relate some worlds, i.e. some worlds might be incomparable. Theorists that are moved by this worry (in particular, Kratzer 1981a) propose the following truth conditions:

(CS4) ⌜If A, would B⌝ is true at w iff, for all A- worlds w''', there is a A-world w'' that is at least as close to w as w''' and such that, for every A-world w' that is at least as close to w as w'', w' is a B-world

These truth conditions require that, for every antecedent-verifying world w', we should be able to find a closer antecedent-verifying world w'', such that every antecedent-verifying world at least as close to w'' verifies the consequent. On these weakened assumptions, $\preceq_w$ induces merely a **preorder** on the set of possible worlds (a preorder is like a total preorder, but admits that some worlds may be incomparable).

For reference, the formal features of comparative similarity on the various accounts are summarized in Table 1.

3.2 Consequences for conditional logic

The accounts in (CS1)–(CS4) are insufficient to fix truth conditions for particular counterfactuals. To do that, more needs to be said about the interpretation of com-

parative similarity. But each of (CS1)–(CS4) is sufficient to fix a logic. This, in turn, leads to predictions about which patterns of inference are valid, and which are not. In this section, I survey some features of the logic of classical truth-conditional accounts. I start by outlining some logical facts that are common to all of (CS1)–(CS4), and then point out some key differences.[6]

3.2.1 Shared logical features of (CS1)–(CS4): nonmonotonicity

All of (CS1)–(CS4) predict that counterfactuals are *nonmonotonic* in the antecedent position. Put simply, this means that adding information to the proposition expressed by the antecedent does not necessarily preserve the truth of a counterfactual. The nonmonotonicity of counterfactuals is illustrated clearly by the failure of the pattern of inference known as Antecedent Strengthening.

Antecedent Strengthening. $A \boxrightarrow C \vDash A^+ \boxrightarrow C \quad$ (with $A^+ \vDash A$)

For a classic counterexample, notice that the inference from (5a) to (5b) is clearly invalid. We can find scenarios where the former is true and the latter false.

(5) a. If I had struck the match, it would have lit.
b. If I had struck the match and the match had been wet, it would have lit.

The failure of antecedent strengthening, and hence the failure of monotonicity, is a natural consequence of an account of counterfactuals based on the notion of minimal change. According to a minimal change account, when we make a counterfactual supposition, we hold fixed as much information as possible from the actual circumstances. If we add extra information to our original supposition, we might contradict some of the information that we previously held fixed. So the failure of Antecedent Strengthening is the logical outcome of the intuitive ideas we started from.

The failure of Antecedent Strengthening is linked to the failure of other, related forms of inference. Here are two that are often mentioned:[7]

Transitivity. $A \boxrightarrow B, B \boxrightarrow C \vDash A \boxrightarrow C$
Contraposition. $A \boxrightarrow C \vDash \neg C \boxrightarrow \neg A$

It's worth emphasizing that, while the accounts in (CS1)–(CS4) invalidate some notable principles, they do validate other principles, some of which correspond to

[6] This section falls far short of a comprehensive survey of counterfactual logic, even focusing only on classical truth-conditional semantics. For an excellent survey of conditional logics, see Egré & Rott (2021).

[7] For counterexamples, see the classical discussions in Stalnaker 1968 and Lewis 1973.

weakened forms of monotonicity. For example, the following, which amounts to Antecedent Strengthening with an added premise, is vindicated by all of (CS1)–(CS4):[8]

Cautious Monotonicity. $A \boxrightarrow B, A \boxrightarrow C \vDash (A \wedge B) \boxrightarrow C$

At first sight, the prediction that Cautious Monotonicity is valid appears vindicated. The inference from (6a) and (6b) to (6c) seems fairly solid. (In ordinary scenarios, one might worry about the truth of (6c), but those worries will apply equally to the truth of (6a).)

(6) a. If I had struck the match, the match would have been wet.
b. If I had struck the match, it would have lit.
c. If I had struck the match and the match had been wet, it would have lit.

At the same time, the validity of Cautious Monotonicity (and related patterns) is not uncontroversial. Bacon 2015, Icard 2017, Santorio 2019 all try to provide counterexamples. So the issue is far from settled.

3.2.2 Logical differences: conditional excluded middle and related patterns

While (CS1)–(CS4) share many features of the logic, they differ in some important ways. Here I will focus on the difference between the logics generated by (CS1) and the logics generated by all the other theories, because it is the one that has received the most attention, and arguably the most theoretically significant.

(CS1), but not the other theories mentioned above, vindicates:

Conditional Excluded Middle. $\vDash (A \boxrightarrow B) \vee (A \boxrightarrow \neg B)$

Conditional Excluded Middle ('CEM' for short) is a conditional counterpart of the principle of Excluded Middle in propositional logic, which states that $\ulcorner A \vee \neg A \urcorner$ is valid.

When comparing his theory to Stalnaker, Lewis declares CEM "the principal virtue and the principal vice of Stalnaker's theory" (1973, p. 79). As Lewis points out, the virtuosity of CEM comes from empirical considerations. One immediate consequence of CEM is that $\neg(A \boxrightarrow B)$ and $A \boxrightarrow \neg B$ are predicted to be equivalent: i.e., bringing negation in and out of the consequent of a counterfactual is predicted to make no difference to meaning. This prediction appears to be confirmed by the facts. For example, the two sentences in (7) appear to be equivalent.

[8]The term 'Cautious Monotonicity' comes from Kraus et al. 1990. The Kraus, Lehmann, and Magidor paper also discusses a number of other interesting principles that are validated by standard counterfactual semantics.

(7) a. It's not the case that, if Ahmed had played Ava, he would have won.
b. If Ahmed had played Ava, he would not have won.

Moreover, instances of CEM sound like tautologies. For example, consider:

(8) If Ahmed had played Ava, he would have won, or, if he had played Ava, he would not have won.

Stalnaker-style theories, like (CS1), predict these data with ease. Conversely, all other theories of counterfactuals struggle with them. For example, on Lewis's theory, it might happen thatm in some of the worlds that the counterfactuals in (8) quantify over Ahmed wins, and in some others he loses. In this case, both the disjuncts in (8) are false, and hence the whole disjunction is false.

Despite these advantages, Lewis suggests that we should abandon (CS1) and adopt instead a semantics in the style of (CS2)–(CS3), on which counterfactuals are universal quantifiers. He gives a theoretical and an empirical argument.

The theoretical argument is that CEM requires an assumption that appears extremely implausible, namely that, for each world *w* and each counterfactual supposition, there should be a unique closest world to *w* that makes that supposition true. Counterexamples are easy to come. Here is a classical case: suppose that I considered flipping a fair coin, but didn't. Is the closest world where I flip the coin a world where the coin lands tails, or one where the coin lands heads? It seems implausible that the question has a determinate answer.

The empirical argument against CEM is that vindicating CEM is in conflict with vindicating another plausible logical principle relating ordinary *would*-counterfactuals with their counterparts involving possibility modals, like (1).

(9) If Ahmed had played Ava, he might have won.

The principle says that *would*-counterfactuals and *might*-counterfactuals are duals:

Duality. $\vDash (A \diamondsuit\!\!\rightarrow B) \leftrightarrow \neg(A \mathrel{\square\!\!\rightarrow} \neg B)$

Duality also enjoys empirical support. Conjunctions of *might*-counterfactuals and of the corresponding *would*-counterfactuals with a negated consequent sound like contradictions:

(10) # If Ahmed had played Ava, he might have won; but if Ahmed had played Ava, he would not have won.

This is immediately predicted if we have Duality. Conversely, accounts like (CS1) struggle with vindicating these data.[9]

[9]For Stalnaker's own take on *might*-counterfactuals, see his discussion in 1984, chapter 7.

Let me highlight the nature of the conflict between CEM and Duality. If we have a classical notion of logical consequence, CEM and Duality jointly entail an unacceptable consequence: *would*- and *might*-counterfactuals are equivalent.[10] So we have three choices: (i) reject CEM; (ii) reject Duality; (iii) reject a classical notion of consequence.

For reasons of space, I cannot discuss the literature on CEM here. But I can offer some pointers to relevant work. Stalnaker (1981; 1984) responds to Lewis's theoretical objection about similarity by appealing to indeterminacy. According to Stalnaker, the correct semantics for counterfactuals is indeed (CS1). But, even though the semantics requires that there is a unique closest antecedent-verifying world for every antecedent, it may be indeterminate which world is the closest one. This indeterminacy can be dealt with using tools imported from the vagueness literature (in particular, supervaluations).

Stalnaker's proposal still requires accepting other model-theoretical constraints, in particular the limit assumption. Swanson (2012) generalizes this approach in an interesting way, proposing to combine Stalnaker's semantics with a version of supervaluationism that does not require the limit assumption.

The appeal to indeterminacy does not address how to vindicate Duality. A few modern accounts take the route of trying to vindicate both CEM and Duality; of course, given the constraints described above, to do this they need to appeal to a nonclassical notion of logical consequence. For example, Schlenker (2004) introduces a trivalent semantics, on which counterfactuals carry a definedness condition, requiring that antecedent worlds be homogenous with respect to the consequent (i.e. that they either all make the consequent true, or all make the consequent false).[11] Santorio (2022a), which discusses indicatives but can be extended to counterfactuals, shows how CEM and Duality can be both vindicated in full by a semantics that combines linear orderings with an informational/dynamic notion of logical

[10]More precisely, we get that $A \Diamond\!\!\rightarrow B$ and $A \Box\!\!\rightarrow B$ are equivalent for every consistent A. The $A \Box\!\!\rightarrow B \vDash A \Diamond\!\!\rightarrow B$ direction of this equivalence follows from standard semantics; here is a proof of the other direction, given CEM and Duality:

i.	$A \Diamond\!\!\rightarrow B$	Assumption
ii.	$\quad A \Box\!\!\rightarrow \neg B$	Supposition for conditional proof
iii.	$\quad A \Box\!\!\rightarrow \neg B \wedge A \Diamond\!\!\rightarrow B$	(i, ii, $\wedge$-Introduction)
iv.	$\quad A \Box\!\!\rightarrow \neg B \wedge \neg(A \Box\!\!\rightarrow \neg B)$	(iii, Duality, substitution of equiv.)
v.	$\quad \bot$	(iv, propositional logic)
vi.	$\neg(A \Box\!\!\rightarrow \neg B)$	(ii-v, Reductio)
vii.	$A \Box\!\!\rightarrow B$	(vi, CEM, Disjunctive syllogism)

[11]See also von Fintel 1997, and Willer 2022 for a conceptually similar attempt in a dynamic framework.

consequence.

3.3 A note on premise semantics

Throughout this section, I have cast standard accounts in the framework of **ordering semantics**, i.e. a family of semantic theories that employ a relation of comparative similarity. But several classical accounts in the literature are stated in a **premise semantics** framework. The basic idea of premise semantics is that conditionals are a kind of enthymematic argument: the consequent is supposed to follow from the antecedent, together with some hidden premises. Premise semantics assigns meanings to counterfactuals appealing to these covert premise sets.

While premise semantics is conceptually and mechanically different from ordering semantics, the results it achieves are similar. In particular, Lewis 1980 proves that premise semantics is equivalent to the version of comparative similarity semantics we stated in (CS4). For this reason, I will only give a quick survey of one kind of premise semantics here, i.e. the premise semantics developed by Kratzer (1977, 1981a, 1981b, 1986, 2012).

The main formal notion of Kratzer's semantics is that of a **conversational background**, which she construes as a function from worlds to sets of propositions. Intuitively, a conversational background models the background information that is used to evaluate a modal claim. Kratzer's official semantics involves two conversational backgrounds, which she calls **modal base** and **ordering source**. For the particular case of counterfactuals, only the ordering source is relevant. Following standard custom, I denote the ordering source with 'g'; hence $g(w)$ is the set of propositions that g yields at w. Kratzer suggests that, to evaluate $\mathsf{A} \boxdot\!\!\rightarrow \mathsf{B}$, we consider all the maximal consistent sets of propositions that (i) include A and (ii) include propositions from the ordering source. $\mathsf{A} \boxdot\!\!\rightarrow \mathsf{B}$ is true just in case B follows from all these sets. More formally:

> ⌜If A, would B⌝ is true with respect to w and g iff, for every maximal consistent set S such that $[\![\mathsf{A}]\!] \in S$ and $S - \{[\![\mathsf{A}]\!]\} \subseteq g(w)$, $S \vDash \mathsf{B}$.

For the particular case of counterfactuals, Kratzer requires that the ordering source should be totally realistic: for each w, $g(w)$ includes all and only the propositions that are true in w (hence, if we construe propositions as sets of worlds: for all w, $\bigcap g(w) = \{w\}$).

4 The problem of similarity

4.1 Similarity and temporal asymmetry

Consider again our starting example:

(1) If I had struck the match, it would have lit.

In §3, I discussed several standard theories of counterfactuals. Without supplementation, all these theories fall short of predicting specific truth conditions for individual counterfactuals like (1). To get predictions of this sort, we need to specify what real-world relation is modeled by the formal similarity relation.

It's important to appreciate that there is a real issue here. Merely appealing to an intuitive notion of similarity, even if appropriately modulated by context, leads to unacceptable results. This point was made by Kit Fine, in his review of Lewis's *Counterfactuals* (1975). Consider:

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

Fine imagines a scenario where then-president Nixon is deciding whether to push the button that launches a nuclear attack against the Soviet Union. In this scenario, (11) is intuitively true. But this is not the result that we get if we appeal to an intuitive, pre-theoretical notion of similarity.

> Now suppose that there never will be a nuclear holocaust. Then that counterfactual is, on Lewis's analysis, very likely false. For given any world in which antecedent and consequent are both true it will be easy to imagine a closer world in which the antecedent is true and the consequent false. For we need only imagine a change that prevents the holocaust but that does not require such a great divergence from reality. (Fine 1975)

Worlds where no nuclear holocaust happens are overall more similar to the actual world than worlds where there is indeed a nuclear holocaust. Hence we appear to predict that the most similar worlds where Nixon pushes the nuclear button are worlds where—perhaps because of a circuit malfunction, perhaps because of some other small-scale event—no nuclear bombs are launched.

We can extract a general moral from Fine's point. The notion of similarity that is relevant for evaluating counterfactuals seems to be crucially sensitive to an asymmetry between past and future. As a first-pass generalization: actual events that are in the past with respect to the event described in the antecedent are relevant to determining similarity, but actual events that are in the future are not. This cannot be overridden by contextual factors or speaker's intentions. To be sure, we can imagine contexts where (11) can be heard as not clearly true, and perhaps even as false. But

the fact that in the actual world a nuclear holocaust has not taken place is irrelevant for evaluating (11).

All classical accounts of similarity try to capture this asymmetry. They agree on a general idea: similarity is determined by laws of nature and matters of particular facts. Given a 'base' world *w*, worlds will be all the more similar to *w* the more they agree with the laws of *w*, and the more they vindicate the particular facts in the history of *w*. But they disagree about which between laws and matters of particular fact should be given up, when we have to accommodate a contrary-to-fact supposition. In particular, we can distinguish two main families of accounts.

Miracles accounts. On some accounts, worlds that are most similar are worlds that share as much of their history with the actual world as possible, and that allow for minor differences in laws of nature. From the perspective of the actual world, these violations of the actual world are 'small miracles', understood as minor deviations from the actual laws. (Crucially, these worlds do not violate their own laws; they merely happen to have different laws from actual laws.)

Lewis (1979) is, famously, a proponent of a miracles account.[12] He proposes a full list of criteria for ranking worlds on the basis of their similarity to the actual world, spelled out as follows:

(i) It is of the first importance to avoid big, widespread, diverse violations of law.

(ii) It is of the second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails.

(iii) It is of the third importance to avoid even small, localized, simple violations of law.

(iv) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

Condition (ii) is the one that ensures that the most similar worlds overlap as much as possible with the actual world in their history. Consider again the Nixon conditional (11). The most similar worlds where Nixon pushes the button are worlds whose history overlaps completely with actual history, up to a time that shortly precedes Nixon's pushing the button. At that point, in those worlds an event that is incompatible with actual laws takes place (perhaps a neuron firing in Nixon's brain). From this point on, history takes an altogether different course.

[12] For other accounts in this family, see Jackson 1977, Bennett 1984, Lange 2000, Kment 2006, Khoo 2022, among others.

No miracles accounts. According to accounts in the second family, the most similar antecedent-worlds are governed by the exact same laws of nature as the actual world. These worlds will involve a divergence in history at some point. Since laws are held fixed, there must be a divergence in history—i.e., a divergence in matters of particular fact—that precedes the antecedent time. On a deterministic picture of laws, this divergence must stretch all the way back to the beginning of history. Accounts of this form are defended, among others, by Nute (1980), Bennett (1984), Loewer (2007), and Dorr (2016).

For illustration, consider again (11). According to no-miracles theories, the most similar worlds where Nixon pushes the button are slightly different from the actual world at every point in history, up to Nixon's pushing the button. Perhaps the difference is microscopic until the event of Nixon's pushing the button. At some point, these slightly different circumstances, in combination with the (actual) laws of nature, bring about Nixon's pushing the button. From here, there are macroscopic divergences in history.

4.2 Similarity and causal dependencies

The literature also contains an altogether different view of similarity. On this view, the notion that is central to similarity is not temporal asymmetry, but rather causal dependency. The most similar worlds to the actual worlds are the ones where causal processes that are independent of the event described by the counterfactual antecedent still unfold as they do in actuality. To put it another way: when evaluating a counterfactual, we 'hold fixed' all the information that is causally independent of the antecedent.

Of course, since there is a temporal asymmetry built into causal dependencies (barring backward causation scenarios), the two views will agree on their verdicts in most cases. But some cases set them apart. One type of cases of this sort are so-called Morgenbesser cases (introduced by Slote 1978 and credited to Sydney Morgenbesser). These cases explore judgments about counterfactuals in scenarios that involve causal dependencies between indeterministic events. Here is a classical example:

> *Coin toss.* Alice is about to toss a coin and offers Bob a bet on heads; Bob declines. Alice tosses the coin, which does indeed land heads.

Consider:

(12) If Bob had taken the bet, he would have won.

(12) seems true, but this judgment cannot be vindicated by theories of similarity based on temporal asymmetry. Suppose that, following Lewis, we let (12) quantify over worlds that perfectly overlap in history with the actual world until shortly before the time at which Bob accepts the bet. Suppose also that coin flips are indeterministic (if you think this is implausible, just modify the example). In the closest worlds where Bob accepts the bet, it is not settled whether the coin lands heads or tails. The coin will land heads in some of them and tails in the others. As a result, (12) is predicted to be not true, contrary to fact.[13]

To get the right prediction, we need to hold fixed not facts that are *in the past* with respect to the antecedent, but rather facts that are *causally independent* of the antecedent. Given the way that the example is constructed, the outcome of the coin flip is causally independent of Bob's decision about the bet.

If we take these cases to be probative, we have a motivation to develop and endorse a causal account of similarity. Accounts of this sort can borrow tools from the causal models framework (see §8), but can still be implemented in a possible worlds semantics that is broadly in the spirit of Lewis and Kratzer (see Kaufmann 2013 and Santorio 2019 for some developed examples).

This concludes my overview of classical theories of counterfactuals. I now move on to consider a number of issues that have emerged in more recent debates.

5 Issues in counterfactual logic

The first set of issues I consider have to do with counterfactual logic. I discuss two points, which turn out to be related.

5.1 Reverse Sobel Sequences and dynamic strict accounts

As we saw in §3, classical accounts of counterfactuals invalidate Antecedent Strengthening. This seems obviously supported by the data. Suppose that I am sitting just next to a large open container filled with water, holding an unstruck match. In this scenario, (13a) does not appear to entail (13b).

(13) a. If I had struck the match, it would have lit.
b. If I had struck the match and had clumsily dropped it in the water right after that, the match would have lit.

[13] The precise verdict will vary depending on which version of comparative similarity semantics we adopt. On (CS1) it is predicted to be indeterminate, and on all of (CS2)-(CS4) it is predicted to be false.

Since Lewis 1973, one classical way to present the data about Antecedent Strengthening involves so-called **Sobel sequences**, i.e. discourses of the form $\ulcorner A \boxright C, (A \wedge B) \boxright C \urcorner$. Here is an example built from the counterfactuals in (13):

(14) If I had struck the match, it would have lit. If I had struck the match and had clumsily dropped it in water right afterwards, the match would not have lit.

Now, the empirical case for the failure of Antecedent Strengthening has been questioned. Notice that, by switching the order of the counterfactuals, we obtain a sequence—a so-called **Reverse Sobel Sequence**—that sounds inconsistent.[14]

(15) # If I had struck the match and had clumsily dropped it in water right afterwards, the match would not have lit. If I had struck the match, it would have lit.

This order-dependence is unexpected on standard semantics. That semantics assigns consistent truth-conditions to (13a) and (13b), so they are predicted to be fine as uttered in any order.

The discovery of Reverse Sobel Sequences spurred the development of a new breed of theories, which build on a dynamic conception of semantics (von Fintel 2001, Gillies 2007, Starr 2014). These theories differ in implementation, but they all share the basic idea. Here I present von Fintel's version. On this semantics, the meaning of counterfactuals is divided into two components. On the one hand, counterfactuals have a 'core' semantic component, which treats counterfactuals simply as universal quantifiers over antecedent worlds. (This semantics is traditionally called **strict conditional** semantics.) On the other, they update, in a way that is systematic and predictable, a parameter that specifies the worlds they quantify over. This parameter, which is called **modal horizon**, tracks what worlds are relevant, against a comparative similarity ordering. The utterance of $A \boxright C$ triggers the expansion of the modal horizon, to include all worlds in the ordering up to the closest A-worlds. Counterfactuals quantify over the expanded modal horizon: $A \boxright C$ says that all the A-worlds in the expanded modal horizon are C-worlds.

Here is a more formal version of the account. Counterfactuals are associated to both truth conditions and to a **context change potential** (CCP). The CCP specifies how a contextual parameter tracking relevant information evolves as a result of the utterance of a counterfactual. Below are statements of both truth conditions and CCP for counterfactuals (I use '*S*' to denote an information state, which is modeled formally as a set of worlds).

Truth conditions

[14] The observation first appeared in print in von Fintel 2001, but it is attributed to Irene Heim.

$[\![\text{A} \boxdot\!\!\rightarrow \text{B}]\!]^{w,f,\preceq}$ is true iff $\forall w' \in f[\text{A} \boxdot\!\!\rightarrow \text{B}]_{w,\preceq} \cap [\![\text{A}]\!]^{w,f,\preceq}$, $w' \in [\![\text{B}]\!]^{w,f,\preceq}$

Context change potential

$f[\text{A} \boxdot\!\!\rightarrow \text{B}]_{w,\preceq} = f(w) \cup \{w' : \forall w'' \in \textsc{max}_{w,\preceq}[\![\text{A}]\!]^{w,f,\preceq}, w' \preceq w''\}$
(where $\textsc{max}_{w,\preceq}[\![\text{A}]\!]^{w,f,\preceq}$ is the set of closest A-worlds to w, relative to $\preceq$)

This semantics is designed to predict order effects. Let's see how it deals with (14) and (15). Suppose that we start with an empty modal horizon: no worlds are relevant. Suppose also that, in the closest worlds where I strike a match, I *don't* drop it in water, and worlds where I strike and drop are further off (Figure 3).

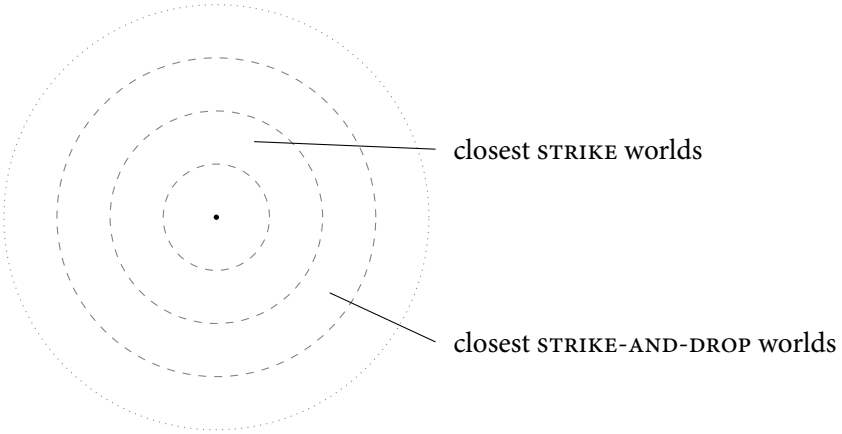


Figure 3: The beginning modal horizon for the match scenario

Now consider sequence (14). (13a), i.e. the first counterfactual in the sequence, expands the modal horizon, letting in the closest worlds where I strike the match. In all those worlds, the match lights up, so the counterfactual is true. (13b), i.e. the second counterfactual, expands the modal horizon further, letting in the closest strike-and-drop worlds (where the match doesn't light). (13b) quantifies over this expanded set of worlds, and is also true (see Figure 4).

Consider now (15). In this case, (13b) appears first, and (13a) second. (13b) expands the modal horizon right away to a set including the closest strike-and-drop worlds. Against this modal horizon, (13b) is true. But now (13a) is also evaluated against this background, and it is false. Hence the sequence is predicted to be infelicitous.

Hence, dynamic strict accounts manage to predict the asymmetry between standard Sobel Sequences and Reverse Sobel Sequences, while preserving the basic insights of nonmonotonic accounts. At the same time, recent literature has pointed

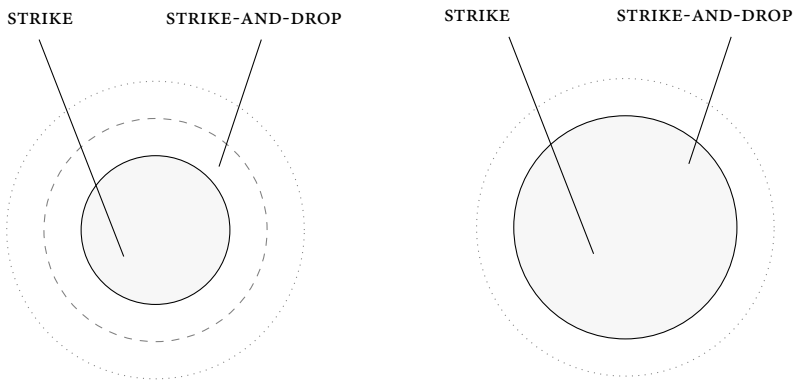


Figure 4: Expansions of the modal horizon triggered by the counterfactuals in (13).

out that these accounts also suffer from drawbacks.[15] Here is one. Dynamic strict accounts predict that all Reverse Sobel Sequences are infelicitous, but it is clear that some of them are perfectly fine (Moss 2012). As an example, consider the following variant on the match scenario.

(16) If I had struck the match but my fingers had been too weak to hold up a match for more than a second, the match would not have lit. But if I had struck the match, it would have lit.

Examples like (16) are a challenge for the dynamic strict account. The latter predicts that Reverse Sobel Sequences are *invariably* infelicitous, since it hardwires the expansion of the modal horizon in the meaning of counterfactuals. So it cannot allow that we have felicitous Reverse Sobel Sequences like (16).[16]

[15]Some other notable problems for the dynamic strict account are pointed out by Nichols 2017. Nichols shows that the expansion of the modal horizon postulated by the dynamic account might let in too many worlds. Notice that the context change potential associated to counterfactuals requires that we let in not just the closest antecent-verifying worlds, but *all* worlds that are at least as close as antecedent verifying worlds. As a result, it may be that, when expanding the modal horizon to let in some $A \wedge B$-worlds, we let in some 'in between' worlds that yield the wrong judgments.

[16]Some further references: see Klecha 2022 for the claim that we can distinguish two kinds of sequences of counterfactuals, one of which is always consistent and reversible; see Lewis 2018 for a contextualist account of the Reverse Sobel Sequences data.

5.2 Import-Export

A second question, which has gained attention more recently, is the validity of Import-Export. Import-Export is the following equivalence principle:

Import-Export. $A \boxrightarrow (B \boxrightarrow C) \dashv\models (A \wedge B) \boxrightarrow C$

Import-Export has been debated extensively in the literature on indicative conditionals. Import-Export appears to be intuitively valid for indicatives, as the perceived equivalence between (17a) and (17b) shows.

(17) a. If the die landed even, then if it didn't land on six, it landed on two.
b. If the die landed even and it didn't land on six, it landed on two.

At first sight, this equivalence seems to hold also for counterfactuals. For example, the sentences in (18) also seem equivalent.

(18) a. If the die had landed even, then if it hadn't landed on six, it would have landed on two.
b. If the die had landed even and it hadn't landed on six, it would have landed on two.

At the same time, the validity of Import-Export leads to clashes with other logical principles. A classical 'collapse' result by Allan Gibbard (1981) shows that any conditional connective that vindicates Import-Export and Modus Ponens, and satisfies plausible side assumptions, has to be equivalent to the material conditional. Recently, Matthew Mandelkern has provided a strengthening of Gibbard's result (2021), showing that the same problem can be generated merely by appealing to Import-Export and Identity, plus side assumptions. (See below for formulations of Modus Ponens and Identity.)

Interestingly, some of the semantic theories that we have discussed validate Import-Export and others don't. In particular, on the classical comparative similarity semantics surveyed in §3, Import-Export fails.[17] Conversely, Import-Export is usually validated by dynamic theories of conditionals.

For current purposes, it's useful to focus on just one fact: given Antecedent Strengthening, we can prove with minimal side assumptions that one of the directions of Import-Export, namely Import, is valid, at least for a large class of conditionals.

[17] To clarify: Import-Export fails on Stalnaker's and Lewis's semantics. The case of Kratzer's semantics is more complicated: since for Kratzer conditionals should not be analyzed as two-place operators, conditionals that involve other conditionals nested in the consequent can be formalized in various ways. As a result, there are several principles that, with some right, might deserve the label 'Import-Export' for Kratzer.

To carry on the derivation, we need the following side principles:

Identity.	$\vDash A \boxrightarrow A$
Modus Ponens.	$A \boxrightarrow B, A \vDash B$
Closure.	If $B_1, \ldots B_n \vDash C$, then $A \boxrightarrow B_1, \ldots A \boxrightarrow B_n \vDash A \boxrightarrow C$

Identity says that "If A, would A" is valid. The validity of Identity is fairly uncontroversial. Modus Ponens is controversial as a general rule[18], but it is taken to be safe in its application to simple conditionals, i.e. conditionals without modal and conditional antecedent and consequent. For current purposes, we can just take the argument to be restricted to this special case. Closure is a principle standardly validated by all counterfactual logics: it says that conditionals preserve entailments in the consequent.[19]

Given these side principles and given Antecedent Strengthening, we derive $(A \wedge B) \boxrightarrow C$ from $A \boxrightarrow (B \boxrightarrow C)$ as follows:

(i)	$A \boxrightarrow (B \boxrightarrow C)$	
(ii)	$(A \wedge B) \boxrightarrow (B \boxrightarrow C)$	(i), Antecedent Strengthening
(iii)	$(A \wedge B) \boxrightarrow (A \wedge B)$	Identity
(iv)	$(A \wedge B) \boxrightarrow B$	(iii), Closure, classical logic
(v)	$(A \wedge B) \boxrightarrow C$	(ii), (iv), MP, Closure

Now, the entailment from Antecedent Strengthening to Import is a potential liability for strict theories, since Import appears to be subject to counterexamples.[20]

To start, let's look at an example presented in Mandelkern (2021) and based on examples by David Etlin and Steve Yablo. Suppose that we have a die (it doesn't matter whether fair or biased); consider:

(19) a. If the die had been thrown and landed four, then if it hadn't landed four, it would have landed two or six.

b. # If the die had been thrown and landed four and it hadn't landed four, it would have landed two or six

(19a) appears to be a consistent and informative claim. (19b) is inconsistent. But the inference from (19a) to (19b) is an instance of Import, hence the pair appears to provide a counterexample to the principle.

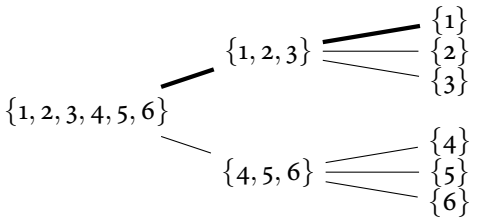
[18] For the *locus classicus* for objection to Modus Ponens for indicative conditionals, see McGee 1985; see also Mandelkern 2020 for an extension of McGee's argument to counterfactuals.

[19] Closure corresponds to the rule "Deduction within Conditionals" in Lewis 1971.

[20] This despite the fact that most instances of both Import and Export appear to be good inferences: counterexamples are quite rare. This is possibly the reason why they have been often overlooked.

A second counterexample is provided by Santorio 2022b. Consider the following scenario:

> **Game show**. A game show involves randomly selecting a number between 1 and 6. To make things more dramatic, the selection works as follows. First, it is determined (via a chancy process) whether the number will be selected among 1, 2, or 3, or among 4, 5, and 6. Then, the final number is selected (again via a chancy process). Suppose that, in actuality, the number 1 is selected (as indicated by the thick line).



Now, consider:

(20) a. If 5 had been selected, then, if an even number had been selected, it would have been 4 or 6.

b. # If 5 had been selected and an even number had been selected, it would have been 4 or 6.

(20a) is informative and, it seems, true, while (20b) appears inconsistent. But the inference from (20a) to (20b) is an instance of Import. So, again, we have an apparent counterexample.

Of course, this discussion is too quick to establish definitive conclusions about the viability of theories of counterfactuals that vindicate Antecedent Strenghtening. But it shows that the debate on Reverse Sobel Sequences has connections with other debates in counterfactual logics. These connections deserve to be explored in greater detail.

6 The role of counterfactual morphology

The second set of issues I discuss concern tense and aspect morphology in counterfactuals. As I will show, questions about morphology are closely related to questions concerning the kind of similarity in play in counterfactuals.

6.1 Arregui's puzzle

Let me start from a puzzle, initially raised by Ana Arregui, Consider the following example (adapted from Arregui 2007):

> *Plants.* I am about to leave for a trip and ask you to come in and water my plants while I'm away. But, just before my trip, my plants die unexpectedly, so I let you know that you don't have to bother after all. You respond:

(21) a. I am sorry, but also a bit relieved. If your plants had died next week (instead), I would be very upset.

b. # I am sorry, but also a bit relieved. If your plants died next week (instead), I would be very upset.

Both (21a) and (21b) involve past tense morphology in the antecedent. In addition, the verb phrase of (21a), but not the one in (21b), involves a perfect. But (21a) is acceptable, while (21b) is not. So, evidently, the difference in morphology produces a difference in acceptability.

The theories of similarity we surveyed in §4 assign no specific role to the morphological features of counterfactuals, and hence are unequipped to predict this difference. For example, on Lewis-style miracle theories, both counterfactuals are predicted to quantify over worlds whose history perfectly matches actual history up to a time that shortly precedes a 'miraculous' turn of events that brings about the antecedent. For the particular case of (21a) and (21b), presumably the divergence occurs shortly before the time of the actual death of the plants. As a result, both conditionals are predicted to quantify over worlds that differ with respect to the actual world with respect to when the speaker's plants die. This prediction is compatible with the felicity of (21a), but not with the infelicity of (21b).

The phenomenon exemplified by (21b) is very general. For further illustration, consider (22)–(24).

(22) Norma arrived yesterday. If she had arrived/# arrived tomorrow instead...

(23) Leslie is sick today. If she had been/#was sick tomorrow instead...

(24) Martin died last month. If he had run/#ran the Boston marathon next month...

Call the problematic counterfactuals 'Arregui counterfactuals'. On a first pass, the problem posed by Arregui counterfactuals is that they appear unable to contradict information that is established in the context. I.e., Arregui counterfactuals cannot be contrary-to-fact. This fact has gone largely undiscussed in the philosophical

literature. But it seems central to our understanding of similarity, and of how counterfactuals select the possibilities they quantify over.[21]

Let me add two clarifications.

First, there is controversy about what features of the antecedent generate the problem. Here I endorse the suggestion in Ippolito 2013, according to which the problematic conditionals have antecedents that involve (i) no perfect and (ii) a future reference time.[22]

Second, it should be clarified in what sense the relevant conditionals are not 'contrary-to-fact'. There are two options. (i) It might be that the antecedent has to be *epistemically possible*, in some sense of epistemic possibility. (ii) It might be that the antecedent has to be *historically possible*. The notion of historical possibility is itself in need of clarification, but the general idea is that it is a metaphysical rather than an epistemic notion: it tracks what possibilities are metaphysically open at a certain time.[23] Both Arregui and Ippolito go for the first option: they require that the antecedents of the problematic *would*-conditionals should be epistemically open, in the sense that they have to be compatible with the common ground of the conversation. But we can show that compatibility with the common ground will not do. Consider the following variant of the plants case:

> *Three plants*. I have three plants at home—a money tree, a fern, and a ficus. While I'm on a trip, I ask my friend to go water my plants. The

[21] Interestingly, counterfactuals of this sort is mentioned in the opening section of Lewis's *Counterfactuals*, only to be set aside:

> [T]here are subjunctive conditionals pertaining to the future, like *If our ground troops entered Laos next year, there would be trouble* that appear to have the truth conditions of indicative conditionals, rather than of the counterfactual conditionals I shall be considering. (p. 4, 1973)

Lewis's attitude might have seemed reasonable at the time of his writing, but is untenable if we want to produce a theory of conditionals that predicts their meaning on the basis of their linguistic form.

[22] This is a departure from the original proposal in Arregui 2007: according to Arregui, the problematic conditionals are those with an antecedent involving (i) no perfect and (ii) an eventive predicate like *die*. Eventive predicates invariably have a future reference time under modals (see Condoravdi 2002 for discussion). So examples like (21b) are insufficient to discriminate between the two hypotheses. Examples like (23), which involve a stative predicate like *be sick*, show that the problem goes beyond eventive predicates, and that it really concerns reference time.

[23] For a characterization of metaphysical openness, see Barnes & Cameron 2009. This characterization appears to take a stance on a number of substantial issues in metaphysics: in particular, it seems to assume that the future is open in a genuinely metaphysical sense. This is not a desirable commitment for a semantics of counterfactuals. But all that I need here is merely that the possibilities quantified over by *would* are different from those quantified over by epistemic modals like *must* and *might*, epistemic conditionals, etc. All that I say is perfectly compatible with the notion of openness being interpreted in a broadly epistemic sense, as long as it's a different notion of epistemic openness from the one used for epistemic modals.

day before I come back, my friend texts me that, despite her best efforts, one of my plants has died, but she doesn't say which.

I love the money tree and the fern, but I don't care much about the ficus. My only reason for keeping it alive is that my mother gave it to me and she's visiting me this weekend. I tell you:

(25) One of my three plants has died, though I don't know which. # If the ficus died next week, I would not mind—I only want my mom to see it alive this weekend.

The *would*-conditional in (25) is still infelicitous. This despite the fact that the antecedent is compatible with the speaker's information in the context: I only know that one of my plants has died, but, for each of my three plants, it is compatible with my information that that plant is alive.

This suggests that Arregui counterfactuals require not that the antecedent be epistemically possible, but rather that the antecedent should be (known to be) historically open: speakers know that the antecedent is not settled, one way or the other, at the time of utterance. To my knowledge, no existing theory of counterfactuals yields this prediction. So Arregui counterfactuals, at least in some respects, are a puzzle for all existing theories.

6.2 Tense and aspect

Arregui counterfactuals show that morphology in counterfactuals matters. What is then, the characteristic morphology of counterfactuals, and what do current theories say about how it impacts the semantics?

Across languages, counterfactuals display some stable morphological features (Iatridou 2000; see Bjorkman & Halpert 2017 for a more nuanced cross-linguistic picture). In a large array of languages (aside from English, Romance languages, modern Greek, Russian, and Polish, to name a few) both the antecedent and the consequent of counterfactuals are in the past tense. Moreover, in many languages (though not all of them—e.g., Russian and Polish are exceptions) counterfactuals involve imperfective aspect.

The presence of past tense is visible in all English examples. Here is one of Iatridou's examples from modern Greek, to illustrate the presence of aspect.

(26) An eperne afto to siropi θa $\gamma_1$inotan kala.
if take/PST/IMPFV this syrup FUT become/PST/IMPFV well
'If he took this syrup, he would get better.' (Iatridou 2000, p. 236)

As Iatridou points out, the tenses and aspect appearing in counterfactuals appear at first sight to be 'fake'. I.e., tense and aspect are not semantically interpreted, or at least they don't appear to be interpreted in the same way they are interpreted outside counterfactuals.

For illustration, consider (26). As for tense: both the antecedent and consequent are in the past tense. Yet the reference time of both of the antecedent and consequent is in the future. Hence, somehow or other, the tense attached to both the antecedent and the consequent does not force them to describe past events. As for aspect: the antecedent and consequent also display imperfective aspect. While the semantics of imperfective aspect is quite complex (see e.g. Ferreira 2016), overall the imperfective represents an event as being in-progress within the boundaries of a relevant temporal interval. Hence, if the imperfective was interpreted in the standard way in counterfactuals, we would expect (26) to mean roughly that the relevant person would be getting better while he was drinking the syrup. But (26) means that the relevant person would get better *after having completed* taking the syrup.

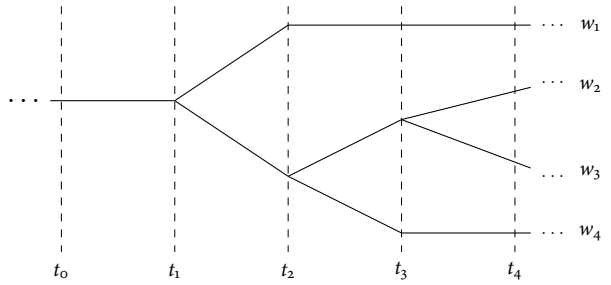
How to interpret this 'fake' morphology? For simplicity, here I will focus on tense alone.[24] There are two main options. The first ('Past-as-Past'; see Arregui 2009 and Ippolito 2013, among others) is to claim that tenses do indeed have their usual meanings in counterfactuals, and give an analysis that explains how they contribute to plausible truth conditions. The second ('Past-as-Modal'; see Iatridou 2000, Schulz 2014a, Mackay 2019, among others) is to claim that tenses are interpreted as having a modal meaning in counterfactuals.

'Past-as-Past' theories consist of a syntactic and a semantic claim. The syntactic claim is that counterfactuals involve a past tense scoping over the main modal of the conditional. Following the literature, here I represent this modal as 'WOLL'. On this view, the logical form of (27a) is in (27b).

(27) a. If Amy was in Rome, Beth would be in Paris.
b. PAST [WOLL[if Amy be in Rome] [Beth be in Paris]]

The semantic claim is nicely illustrated via branching worlds diagrams. Worlds are represented as lines. When two worlds have identical histories, up to a point, the relevant lines overlap; when they diverge in history, their lines branch away from each other.

[24]Much of the literature has taken this route, claiming that aspect does not have a crucial role to play. See e.g. Iatridou 2000 for discussion. Here I set aside aspect not because I endorse this position, but just for simplicity.



The idea behind Past-as-Past theories is that the past tense in *would* shifts back the point at which we evaluate the rest of the sentence to some past time in the diagram. Hence, roughly, *would* [*if* [A, B]] means that there is a past time at which *will* [*if* [A, B]] holds.

'Past-as-Modal' theories start from the idea that, in some linguistic contexts, past tense morphology receives a modal interpretation. One way of formulating this idea is that the past tense has an unspecific meaning. Abstractly, this meaning can be characterized as conveying distance or difference from a designated element. For example, following Iatridou (2000), we can take the past tense to mean that the entity that is topically relevant—roughly, the entity 'talked about'—does not overlap with the entity that is contextually relevant:

$$\text{Topic}(x) \text{ excludes } \text{Context}(x)$$

The entities in question can be times or sets of worlds. When they are times, the past tense gets its usual temporal meaning: the relevant time (i.e., the reference time of the clause) excludes the contextually salient time, i.e. the present time. When they are sets of worlds, the past tense gets a modal meaning: the relevant set of worlds (i.e. the worlds that are 'talked about' by a modal sentence) excludes the contextually salient world, i.e. the actual world.

To end, let me emphasize one moral: morphology matters for the semantics of counterfactuals. As Arregui's puzzle shows, subtle differences in tense morphology can make a difference to what worlds counterfactuals quantify over, and hence to truth conditions. A full theory of comparative similarity in counterfactuals passes through a study of how tense, aspect, and possibly other morphological features, contribute to determining what worlds count as 'most similar'.

7 The relation between counterfactuals and probability

The third set of issues that I take up concerns the relation between counterfactuals and probability. Suppose that I consider buying a ticket in a lottery with a million

tickets, but eventually don't.[25] Now take:

(28) If I had bought a lottery ticket, I would have lost.

What credence should we assign to (28)? Most people who are unbiased by prior theoretical commitments would say that we should assign it high credence. This seemingly simple answer raises a number of theoretical issues.

First, several of the classical theories in §3 have difficulties accommodating it. In particular, on plausible assumptions about similarity, all of (CS2)–(CS4) predict that (28) is false, and that hence should get zero, or near-zero probability. So, if we take seriously questions concerning credence in counterfactuals, we appear to have arguments against some classical accounts, and in favor of the Stalnaker-style semantics in (CS1).[26]

More generally, we might expect that intuitions about probabilities of counterfactuals should be systematized. We should find some general bridge principle that specifies what probability one should assign to counterfactuals, as a function of the probabilities of the antecedent and the consequent. This expectation is bolstered if we consider variants of the lottery case. Suppose that we vary the number of tickets that are sold in the lottery. If we make the number larger, the intuitive probability of (28) goes up. Conversely, if we make it lower, it goes down. So it seems that there is a systematic connection between the probability of a counterfactual and the probability of its antecedent and consequent.

The benchmark proposal to capture this connection is due to Skyrms 1980 and is the following: your credence in $\mathsf{A} \boxrightarrow \mathsf{B}$ should equal your expectation of the past chance of B, given A. Let me state this formally. Let a chance function be a function $ch_{w,t}(\cdot)$ that assigns probability to a propositon, relative to a world w and a time t.[27] The principle defended by Skyrms is:

> **Skyrms' Thesis**
> For all A, B, and for all rational credence functions cr_{E_t} such that E_t is the subject's total evidence at t:

[25] This argument is a variant on the lead example in Schulz 2014b, to whom I am indebted also for some of the points below; see also Schulz 2017. A version of the argument appears already in Edgington 2008.

[26] As a historical note: judgments about probability were exactly one of the arguments that Stalnaker initially used to support his own theory (1970) against Lewis's theory. The argument was blunted by Lewis's discovery of triviality results (1976), but of course the intuitions about probability still need to be accounted for.

[27] For a more explicit, and in my view better, way of construing chance functions, see Meacham 2010.

$$cr_{E_t}(\mathsf{A} \mathrel{\square\!\!\rightarrow} \mathsf{B}) = \sum_{w \in W} cr_{E_t}(w) \times ch_{w,t^-}(\mathsf{B} \mid \mathsf{A})$$

Notice that Skyrms' Thesis involves a shifted time-index t^- on the chance function: we should consider not the current chances, but rather the chances that obtained at some point in the past (the time 'just before' the truth status of the counterfactual antecedent was settled).

For illustration, suppose that Lily considered flipping a coin yesterday at 1pm, but in the end she didn't. Consider:

(29) If Lily had flipped the coin, it would have landed heads.

We are certain (suppose) that the coin is fair. So we are certain that, at the relevant time (i.e. a time 'just before' Lily decided not to flip) the chance of heads conditional on flipping was 1/2. As a result, in this case Skyrms' Thesis gives us that our rational credence in (29) should be 1/2.

Skyrms' Thesis has several upsides. To start, it gets a number of intuitions right, including those about lottery cases discussed above. Moreover, the connection between rational credence in counterfactuals and chance seems clearly correct. The credences we assign to counterfactuals appear to track an element that is not purely epistemic, and chance seems to play the right role.[28] At the same time, Skyrms' Thesis, on a par with similar bridge principles concerning the probabilities of indicative conditionals, is subject to so-called triviality results.

Roughly, triviality results (so-called after the initial result of this sort, proved in Lewis 1976) are results showing that bridge principles like Skyrms' Thesis, in combination with some simple assumptions, lead to unacceptable conclusions. Lewis famously showed that, if we endorse the intuitive idea that the rational credence in an indicative conditional *If A, B* should equal the conditional probability of B given A, we can prove that the probability of *If A, B* should equal the probability of B, for any A and B—an absurd result. Similarly, Williams 2012 shows that we can prove an analogous result from Skyrms' Thesis. Given minimal side assumptions, we can prove that the probability of a counterfactual should equal the probability of its consequent. Again, this is an absurd result, as is shown by the pair below (assume that the die in question is a regular 6-sided die).

[28] Indeed, as Santorio 2022b points out, if we have the Principal Principle we can show that Skyrms' Thesis is equivalent to a principle that is entirely about chance, namely:

Chancy Equation $\quad ch_{w,t}(\mathsf{A} \mathrel{\square\!\!\rightarrow} \mathsf{B}) = ch_{w,t^-}(\mathsf{B} \mid \mathsf{A})$

Informally, the Chancy Equation says that the chance of a counterfactual A $\mathrel{\square\!\!\rightarrow}$ B (at w, t) equals the conditional chance of B given A (at w, t^-).

(30) a. If the die landed even, it would land on 4.

b. The die will land on 4.

The literature on this topic, which is rapidly developing, contains a number of accounts that try to replace Skyrms' Thesis with a different, triviality-free principle. Some attempts involve endorsing restricted versions of Skyrms' Thesis (see e.g. Khoo 2022), replacing chance with an epistemic notion of probability (Schultheis 2022), or rejecting the idea that conditional chances capture, in general, the way that we should update the chance function for the purposes of formulating a bridge principle between counterfactuals and probability (Santorio 2022b).

8 Counterfactuals and causal models

The last issue I want to consider is the connection between counterfactuals and causal models. The causal models framework is a formal framework used to model both real causal connections and causal reasoning (Galles & Pearl 1998, Halpern 2000, Pearl 2000, Spirtes et al. 2000 among many). The framework can be used to evaluate counterfactuals, which in the original framework are thought of as capturing external changes ('interventions') on a model. In recent years, much work in philosophical logic and formal semantics has tried to capture the ideas behind the causal model frameworks and implement them in a semantics for counterfactuals. (See, among many, Briggs 2012, Kaufmann 2013, Icard 2017, Ciardelli et al. 2018, Santorio 2019). Here I provide a survey of how causal models deal with counterfactuals, and raise some open issues. The introduction is very informal and leaves out applications of the framework in a probabilistic setting. But it is sufficient to capture the main ideas.

8.1 The framework and the evaluation of counterfactuals

A causal model consists of two elements: a set of **random variables**, and a set of **structural equations**. A random variable can be thought of as a set of mutually exclusive and jointly exhaustive outcomes for a process. In philosophy and semantics, this structure represents a partition of logical space, and is often used to capture the denotation of an interrogative clause (see, among many, Lewis 1982, and Lewis 1988, Groenendijk & Stokhof 1984). Hence we may think of a random variable as having the content of a question. For current purposes, I assume that random variables can have the values 1 and 0 (1 for the case in which the relevant event happens, 0 for the case in which it doesn't). Structural equations are mathematical equations that state the relations between different values of random variables.

Let me go through an example in detail. Consider again the case of a dry match that has not been struck. For simplicity, suppose that the only factors that have any bearing on the match lighting are whether it is struck, and whether it is dry. This is a causal model for this simple scenario.

Random variables	**Structural equations**
S: whether the match is struck D: whether the match is dry L: whether the match lights	L = min(S, D)

Random variables are traditionally divided into *exogenous* and *endogenous* ones. Exogenous variables are those whose values are determined by factors external to the model. Endogenous variables, conversely, are those whose values are determined by factors within the model. In this model, L is the only endogenous variable.

Causal models are usually represented visually by means of directed graphs, i.e. diagrams in which nodes represent random variables and arrows represent relationships of causal dependence. This is the graph corresponding to our toy model:[29]

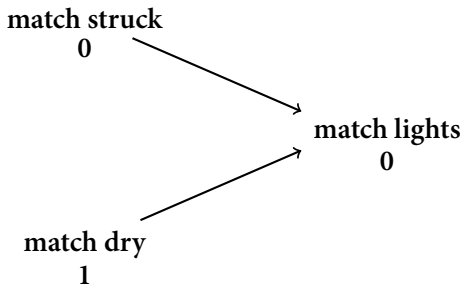


Figure 5: A simple causal model for the match scenario.

As the figure shows, the values of the exogenous variables are: 0 for S, and 1 for D. From here, via the structural equation we determine that the value of L is 0.

[29]Notice that the visual representation generally produces a loss of information. The arrows represent causal dependence, but they are silent about exactly what that dependence involves. For example, the graph above doesn't specify whether the dependence between L and the other two variables is conjunctive or disjunctive (i.e. the same graph is compatible with the equation L = max(S, D)). So graphs are really just convenient props; the full specification of a causal model is given by the set of variables and structural equations.

How can we use causal models to evaluate counterfactuals? The key notion is that of an *intervention*. An intervention consists in manipulating one of the variables 'from the outside', i.e. without adjusting the values of the causal variables that are upstream. Formally, making an intervention amounts to changing the initial model. We strike out the equation that determines the value of the variable that is intervened on, and we replace it with a new equation that simply states the value of that variable. Then we recalculate all the values of the variables that are downstream. We then use the new model to assess the consequent.

Consider again the match scenario. Suppose we want to evaluate the Goodman-style counterfactuals:

(31) a. If the match had been struck, it would have lit.

b. If the match had been struck, it would have been wet.

Since they have the same antecedent, (31a) and (31b) generate the same intervention: they change the initial value of the exogenous variable S from 0 to 1. Here is a graph of the new model:[30]

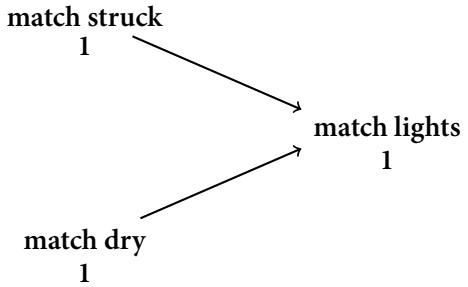


Figure 6: The causal model we obtain after an intervention that sets the value of S to 1.

As you can read from the values of the variables, this procedure predicts that (31a) is true and (31b) is false, as desired. So using models that explicitly encode information about causal dependencies is helpful in explaining the asymmetry between the two counterfactuals.

[30]Notice that, in this particular case, the only structural equation S = min(L, D) remains in place, since the variable intervened on is exogenous. But, in general, interventions will lead to replacing structural equations.

8.2 Problems and prospects for causal models in semantics

Encoding causal dependencies in the semantics for counterfactuals is helpful for a number of purposes. In addition to the Goodman problem described above, a causal models-inspired framework yields correct predictions in the Morgenbesser-type scenarios of §4.2 (see e.g. Hiddleston 2005 for discussion). More generally, there seems to be something both intuitive and powerful about using causal structures to evaluate counterfactual reasoning. So it's unsurprising that a lot of recent work in philosophy of language and semantics has gone in this direction.

Of course, the procedure for evaluating counterfactuals sketched above falls far short of a compositional semantics.[31] But interventionist reasoning can be imported into more standard semantic frameworks, such as a truthmaker semantics (as shown by Briggs 2012), or a Kratzer-style premise semantics, by adding some extra structure to premise sets (see e.g. Kaufmann 2013) or to the premises themselves (see Santorio 2019). All these approaches incorporate at least some of the intuitions behind the causal models framework, and produce interesting variants of classical semantics and logics for counterfactuals.

As it is, however, one major problem remains unsolved. The predictions of causal models-type reasoning strongly depend on the choice of model, i.e. on what variables and structural equations we decide to use. So, to specify a semantics for natural language counterfactuals in terms of causal models, we have to specify how a context of utterance fixes a choice of variables and equations. This is the place where the metasemantic questions concerning the linking of formal parameters to intuitive notions re-emerge. Simply switching to a causal models-inspired semantics provides no easy fix.

9 Conclusion

Counterfactuals are at the crossroads of some of the hardest questions in philosophy of language, as well as philosophy in general. On the language side alone, counterfactuals have links to theories of modality, tense, aspect, and the connection between modality and probability. More generally, counterfactuals are of central relevance for theories of causation, explanation, and laws of nature, and are used throughout philosophical debates. This overview has been obviously partial. But

[31] Just to highlight a few issues: it cannot handle complex antecedents; it is not integrated with compositional treatment of other phenomena related to counterfactuals, such as tense and aspect; it does not involve an explicit statement of what contextual parameters are involved in the evaluation of counterfactuals.

hopefully it shows that, despite decades of work on the topic, there are a number of open problems that deserve attention. Much work is still needed to get counterfactuals right.

References

Anderson, Alan Ross (1951). "A note on Subjunctive and Counterfactual Conditionals." *Analysis*, 12(2): pp. 35–38.

Arregui, Ana (2007). "When Aspect Matters: The Case of Would-Conditionals." *Natural Language Semantics*, 15(3): pp. 221–264.

Arregui, Ana (2009). "On Similarity in Counterfactuals." *Linguistics and Philosophy*, 32(3): pp. 245–278.

Bacon, Andrew (2015). "Stalnaker's Thesis in Context." *Review of Symbolic Logic*, 8(1): pp. 131–163.

Barnes, Elizabeth, and Ross Cameron (2009). "The open future: Bivalence, determinism and ontology." *Philosophical Studies*, 146(2): pp. 291–309.

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

Bjorkman, Bronwyn M, and Claire Halpert (2017). "In an imperfect world: Deriving the typology of counterfactual marking." *Modality Across Syntactic Categories*, 63: pp. 157.

Briggs, Rachael (2012). "Interventionist Counterfactuals." *Philosophical studies*, 160(1): pp. 139–166.

Ciardelli, Ivano, Linmin Zhang, and Lucas Champollion (2018). "Two switches in the theory of counterfactuals: A study of truth conditionality and minimal change." *Linguistic and Philosophy*, 41: pp. 577–621.

Condoravdi, Cleo (2002). "Temporal interpretation of modals-modals for the present and for the past." In D. I. Beaver, L. D. C. Martinez, and B. Z. Clark (eds.) *The construction of meaning*, pp. 59–88.

Dorr, Cian (2016). "Against Counterfactual Miracles." *Philosophical Review*, 125(2): pp. 241–286.

Edgington, Dorothy (2008). "Counterfactuals." *Proceedings of the Aristotelian Society*, 108: pp. 1–21.

Egré, Paul, and Hans Rott (2021). "The Logic of Conditionals." In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. Winter 2021 ed.

Ferreira, Marcelo (2016). "The semantic ingredients of imperfectivity in progressives, habituals, and counterfactuals." *Natural Language Semantics*, 24(4): pp. 353–397.

Fine, Kit (1975). "Review of Lewis' Counterfactuals." *Mind*, 84: pp. 451–458.

von Fintel, Kai (1997). "Bare plurals, bare conditionals, and only." *Journal of Semantics*, 14(1): pp. 1–56.

von Fintel, Kai (2001). "Counterfactuals in a Dynamic Context." *Current Studies in Linguistics Series*, 36: pp. 123–152.

von Fintel, Kai, and Sabine Iatridou (2022). "Prolegomena to a Theory of X-Marking." Unpublished draft, available at `http://web.mit.edu/fintel/fintel-iatridou-2022-x.pdf`.

Galles, David, and Judea Pearl (1998). "An axiomatic characterization of causal counterfactuals." *Foundations of Science*, 3(1): pp. 151–182.

Gibbard, Allan (1981). "Two Recent Theories of Conditionals." In W. Harper, R. C. Stalnaker, and G. Pearce (eds.) *Ifs*, Reidel, pp. 211–247.

Gillies, Anthony S (2007). "Counterfactual scorekeeping." *Linguistics and Philosophy*, 30(3): pp. 329–360.

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

Goodman, Nelson (1955). *Fact, Fiction, and Forecast*. Harvard University Press.

Groenendijk, Jerome, and Martin Stokhof (1984). *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Halpern, Joseph (2000). "Axiomatizing Causal Reasoning." *Journal of Artificial Intelligence Research*, 12: pp. 317–337.

Hiddleston, Eric (2005). "A Causal Theory of Counterfactuals." *Noûs*, 39(4): pp. 632–657.

Iatridou, Sabine (2000). "The grammatical ingredients of counterfactuality." *Linguistic inquiry*, 31(2): pp. 231–270.

Icard, Thomas (2017). "From Programs to Causal Models." In *Proceedings of the 21st Amsterdam Colloquium*, pp. 35–44. University of Amsterdam.

Ippolito, Michela (2013). *Subjunctive conditionals: A linguistic analysis*, vol. 65. MIT Press.

Jackson, Frank (1977). "A causal theory of counterfactuals." *Australasian Journal of Philosophy*, 55(1): pp. 3–21.

Kaufmann, Stefan (2013). "Causal Premise Semantics." *Cognitive science*, 37(6): pp. 1136–1170.

Khoo, Justin (2022). *The Meaning of "If"*. New York, USA: Oxford University Press.

Klecha, Peter (2022). "On the Consistency and Reversibility of Certain Sequences of Counterfactual Assertions." *Mind*, 131(521): pp. 1–33.

Kment, Boris (2006). "Counterfactuals and Explanation." *Mind*, 115(458): pp. 261–310.

Kratzer, Angelika (1977). "What 'Must' and 'Can' Must and Can Mean." *Linguistics and Philosophy*, 1(3): pp. 337–355.

Kratzer, Angelika (1981a). "The Notional Category of Modality." In H. J. Eikmeyer, and H. Rieser (eds.) *Words, Worlds, and Contexts: New Approaches to Word Semantics*, Berlin: de Gruyter.

Kratzer, Angelika (1981b). "Partition and Revision: The Semantics of Counterfactuals." *Journal of Philosophical Logic*, 10(2): pp. 201–216.

Kratzer, Angelika (1986). "Conditionals." In *Chicago Linguistics Society: Papers from the Parasession on Pragmatics and Grammatical Theory*, vol. 22, pp. 1–15. University of Chicago, Chicago IL: Chicago Linguistic Society.

Kratzer, Angelika (2012). *Modals and Conditionals: New and Revised Perspectives*, vol. 36. Oxford University Press.

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor (1990). "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics." *Artificial intelligence*, 44(1): pp. 167–207.

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

Lewis, David (1971). "Completeness and Decidability of Three Logics of Counterfactual Conditionals." *Theoria*, 37(1): pp. 74–85.

Lewis, David (1976). "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review*, 85(3): pp. 297–315.

Lewis, David (1980). "A subjectivist's guide to objective chance." In *Ifs*, Springer, pp. 267–297.

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

Lewis, David K. (1979). "Counterfactual dependence and time's arrow." *Noûs*, 13(4): pp. 455–476.

Lewis, David K. (1982). "Logic for Equivocators." *Noûs*, 16(3): pp. 431–441.

Lewis, David K. (1988). "Relevant Implication." *Theoria*, 54(3): pp. 161–174.

Lewis, Karen S. (2018). "Counterfactual Discourse in Context." *Noûs*, 52(3): pp. 481–507.

Loewer, Barry (2007). "Counterfactuals and the Second Law." In H. Price, and R. Corry (eds.) *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press.

Mackay, John (2019). "Modal interpretation of tense in subjunctive conditionals." *Semantics & Pragmatics*, 12(2): pp. 1–29.

Mandelkern, Matthew (2020). "A Counterexample to Modus Ponenses." *Journal of Philosophy*, 117(6): pp. 315–331.

Mandelkern, Matthew (2021). "If P, Then P!" *Journal of Philosophy*, 118(12): pp. 645–679.

McGee, Vann (1985). "A Counterexample to Modus Ponens." *Journal of Philosophy*, 82(9): pp. 462–471.

Meacham, Christopher J. G. (2010). "Two Mistakes Regarding the Principal Principle." *British Journal for the Philosophy of Science*, 61(2): pp. 407–431.

Moss, Sarah (2012). "On the Pragmatics of Counterfactuals." *Noûs*, 46(3): pp. 561–586.

Nichols, Cory (2017). "Strict Conditional Accounts of Counterfactuals." *Linguistics and Philosophy*, 40(6): pp. 621–645.

Nute, Donald (1980). "Conversational scorekeeping and conditionals." *Journal of Philosophical Logic*, 9(2): pp. 153–166.

Pearl, Judea (2000). *Causality: models, reasoning and inference*. Cambridge University Press.

Santorio, Paolo (2019). "Interventions in Premise Semantics." *Philosophers' Imprint*, 19(1).

Santorio, Paolo (2022a). "Path semantics for indicative conditionals." *Mind*, 131(521): pp. 59–98.

Santorio, Paolo (2022b). "Probabilities of Counterfactuals are Counterfactual Probabilities." Draft, University of Maryland, College Park.

Schlenker, Philippe (2004). "Conditionals as definite descriptions." *Research on language and computation*, 2(3): pp. 417–462.

Schultheis, Ginger (2022). "Counterfactual Probability." Forthcoming in *Journal of Philosophy*.

Schulz, Katrin (2014a). "Fake tense in conditional sentences: A modal approach." *Natural language semantics*, 22(2): pp. 117–144.

Schulz, Moritz (2014b). "Counterfactuals and Arbitrariness." *Mind*, 123(492): pp. 1021–1055.

Schulz, Moritz (2017). *Counterfactuals and probability*. Oxford University Press.

Skyrms, Brian (1980). *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. Yale University Press.

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

Spirtes, Peter, Clark Glymour, Scheines N., and Richard (2000). *Causation, Prediction, and Search, 2nd edition*. Cambridge, MA: MIT Press.

Stalnaker, Robert (1968). "A Theory of Conditionals." In N. Recher (ed.) *Studies in Logical Theory*, Oxford.

Stalnaker, Robert (1981). "A Defense of Conditional Excluded Middle." In W. Harper, R. C. Stalnaker, and G. Pearce (eds.) *Ifs*, Reidel, pp. 87–104.

Stalnaker, Robert (1984). *Inquiry*. Cambridge University Press.

Stalnaker, Robert C. (1970). "Probability and Conditionals." *Philosophy of Science*, 37(1): pp. 64–80.

Starr, William (2014). "What *if*?" *Philosophers' Imprint*, 14(10).

Swanson, Eric (2012). "Conditional Excluded Middle Without the Limit Assumption." *Philosophy and Phenomenological Research*, 85(2): pp. 301–321.

Willer, Malter (2022). "Negating Conditionals." In *Oxford Studies in Philosophy of Language, Volume II*, Oxford University Press, p. 234–266.

Williams, J. Robert G. (2012). "Counterfactual Triviality." *Philosophy and Phenomenological Research*, 85(3): pp. 648–670.